Computational approaches to the analysis of the T cell receptor repertoire

Katharine Best

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Division of Infection & Immunity

University College London

I, Katharine Best, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

The T cell receptor (TCR) repertoire has the potential to be a highly personalised biomarker of historic or current immune challenges, and may hold clinically relevant information. This thesis reviews aspects of the measurement and analysis of the TCR repertoire, including approaches to obtaining high-throughput sequencing data and using these data to investigate features of the repertoire in health and disease. The thesis then considers three topics related to computational and experimental analysis of the TCR repertoire.

First, this thesis explores a technical challenge in obtaining accurate quantitative TCR repertoire sequence data, observing substantial heterogeneity in the PCR amplification step essential for most current high-throughput sequencing protocols. An important conclusion of this chapter is that single molecule barcoding before amplification is essential to obtain robust quantification of clone abundances from sequence data.

The second chapter considers the challenges of producing an effective TCR repertoire which can provide broad coverage of potential pathogens while maintaining tolerance to self-peptides. A computational model is explored which incorporates a linear programming representation of peripheral tolerance, with dendritic cells acting as the central agents reshaping the T cell population. The model is shown to maintain a population with restricted responsiveness to self-peptides while retaining a diverse and cross-reactive repertoire.

In the final results chapter, TCR repertoire data from immunised mice is used to demonstrate that within a simplified animal model of immune response, the antigen responsive CDR3βs are almost completely private. However, exploration of the protein sequences of the antigen associated CDR3βs suggests that there may be amino acid motifs defining the antigen response.

Overall, this thesis demonstrates the application of computational and modelling approaches to address questions regarding the TCR repertoire, facilitating interpretation of high-throughput sequencing data and providing insight into maintenance of diversity in the peripheral T cell population.

Acknowledgements

Firstly, thanks must go to my primary supervisor, Prof. Benny Chain, for his continued support and encouragement without which this thesis wouldn't exist. Thanks also go to my secondary supervisor, Prof. John Shawe-Taylor and to Dr. Maddy Noursadeghi who have both provided invaluable input and advice during my PhD.

I am particularly grateful to Dr. Chris Watkins, with whom the model presented in Chapter 3 was developed, and to Prof. Nir Friedman and his group who generously provided the data analysed in Chapter 4.

Thanks go to everyone in the Chain and Noursadeghi lab groups, for their help and particularly for their patience in the face of my frequent questions.

The support of my family and friends has also been essential throughout this time and I'm immensely grateful to all of them.

Contents

List of Figures

List of Tables

List of abbreviations

APC	antigen presenting cell
ART	anti-retroviral therapy
CDR	complementarity determining region
CFA	complete Freund's adjuvant
CMV	cytomegalovirus
CTCL	cutaneous T cell lymphoma
DC	dendritic cell
DCR	five-part Decombinator classifier describing a TCR
gDNA	genomic DNA
GvHD	graft vs host disease
HIV	human immunodeficiency virus
HSCT	haematopoietic stem cell transplantation
HTS	high throughput sequencing
Ig	immunoglobulin
MHC	major histocompatibility complex
MRD	minimal residual disease
mRNA	messenger RNA
OVA	ovalbumin protein or peptide
PBS	phosphate buffered saline
PCR	polymerase chain reaction
pMHC	peptide/MHC complex
RSS	recombination signal sequences
SI	Shannon information
spMHC	self-peptide/MHC complex
SRD	short read Decombinator
T-ALL	acute T lymphoblastic leukemia
TB	mycobacterium tuberculosis
TCR	T cell receptor
TIL	tumour infiltrating lymphocyte
TNF	tumour necrosis factor

Chapter 1

Introduction

1.1 The TCR repertoire

The immune system in an individual needs to protect against harmful pathogens in order to avoid disease. The cells of the innate immune system (including monocytes, macrophages, neutrophils and dendritic cells) each act to enable rapid detection of the presence of pathogens and to launch broad immune responses. Additionally, certain cells of the innate immune system are able to initiate an adaptive immune response via activation of T and B lymphocytes in an antigen-specific manner. The antigen-specificity of the T and B lymphocytes is determined by their T or B cell receptors, and the repertoire of antigen-specific receptors in the lymphocyte population of an individual provides information about the status of their immune system.

1.1.1 TCR recognition of antigen

T cells recognise antigen through interactions between the T cell receptor ('TCR') and peptide fragments in the context of major histocompatibility complex ('MHC') on the surface of an antigen presenting cell ('APC') such as a dendritic cell. There are a number of factors that influence whether a particular TCR is able to recognise and respond to a particular peptide-MHC ('pMHC') [82], and here the recognition is discussed more generally.

The TCR is a membrane-bound heterodimer and in most T cells is formed of an α

and a β chain, although a minority of T cells express a receptor formed of a γ and a δ chain. Each chain contains constant ('C') and variable ('V') regions (Figure 1.1a). The variable portion of the TCR is formed by imprecise rearrangement of gene segments, discussed in more detail in Section 1.1.2. There are a number of areas of the TCR that are responsible for specific recognition of pMHC, each known as a complementarity determining region ('CDR'). Each chain of the TCR has a highly variable region called the CDR3, which is defined as the section between two conserved amino acid motifs. The CDR3 is highly variable due to non-templated nucleotide additions and nucleotide deletions occuring in this region of the TCR during the rearrangment process. The CDR3 is the part of the TCR which is primarily responsible for contact with, and therefore recognition of, the pMHC and as such is the major part of the TCR defining the antigen specificity of the T cell [71]. Other regions of the receptor are mostly responsible for contact with parts of the MHC, ensuring a stable and productive interaction between the T cell and the APC.

MHC molecules are glycoproteins and come in two different classes, class I and class II. The classes are expressed to different degrees in different cell types, and have different structures. However, both are formed of 4 subunits, with the pair of subunits furthest from the membrane forming a 'groove' into which a peptide fragment can bind (Figure 1.1b, c). Polymorphisms in both classes of MHC molecules are mostly located in the peptide-binding cleft [20], meaning that individuals with different MHC types will be able to present different peptide fragments to the T cell population.

APCs, such as dendritic cells, process proteins into short peptide fragments that may be able to be presented in groove of MHC molecules and these pMHC complexes are then transported to the surface of the cell [150]. The proteins that are processed and presented can either be cellular or exogenous, in which case they must be endocytosed or phagocytosed into the antigen presenting cells. Only peptide stretches that are compatible with the MHC molecule are able to be presented, and as such both the overall structure of the MHC and the MHC type of the individual limits the antigens presented to T cells. In general class I MHC can only present a more restricted set of peptides, with the majority of presented peptides being 9 or 10 amino acids long, while class II can accomodate longer peptides [62].

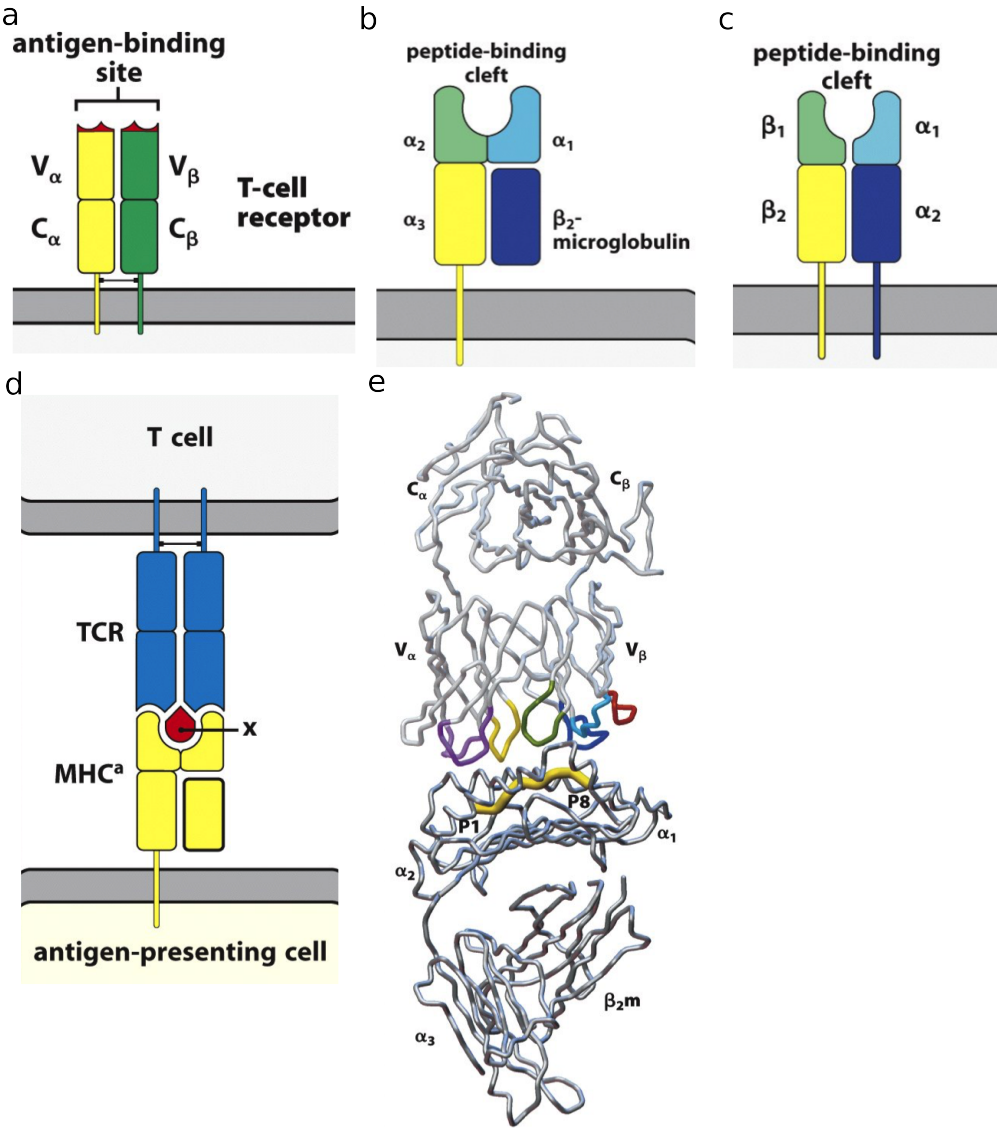

Figure 1.1: TCR structure and recognition of peptide-MHC complexes
All panels taken from [107]
(a) The T cell receptor is formed of two chains (generally α and β, although some T cells express a receptor formed of γ and δ chains), each consisting of a constant region (C) and a variable region (V). The region of the TCR responsible for antigen recognition spans the variable regions of both chains.
(b) MHC Class I molecules are formed of a membrane bound α chain, consisting of three domains (α_1, α_2 and α_3) and a β_2 microglobulin.
(c) MHC Class II molecules are formed of two membrane-bound chains (α and β), each consisting of two domains.
(d) When a TCR expressed on a T cell interacts with sufficient binding strength with a compatible peptide fragment (x) in the context of a compatible MHC molecule presented by an antigen presenting cell there is antigen recognition.
(e) When a TCR binds to a peptide:MHC class I complex the TCR interacts with parts of both the α_1 and α_2 domains of the MHC and with parts of the presented peptide fragment.

With some exceptions (for example, invariant T cells which recognise non-MHC antigen, and activation by super-antigens which does not require costimulation), T cells are activated through interaction between TCR and pMHC on the surface of an APC (Figure 1.1d,e), in the presence of costimulatory signals and appropriate environmental conditions. In order for a TCR-pMHC interaction to yield T cell activation and an immune response there needs to be sufficient affinity between the receptor and the antigen. Understanding the factors that determine whether a TCR will interact with a pMHC with sufficient affinity or for a sufficient duration to initiate activation is therefore a key question if we wish to understand whether a given T cell population will mount a succesful immune response against a given pathogen, or if we want to know which clones in a repertoire are important in a particular disease setting. However, how to predict the interaction between the TCR and pMHC, and therefore produce a mapping between TCR clones and the antigens to which they are able to mount an immune response, is not fully understood.

1.1.2 Generation of a diverse TCR repertoire

In order to be able to recognise and respond to the unknown potential pathogens an individual might encounter, the T cell population needs to express a range of TCRs with diverse specificities. This is achieved by stochastic recombination of gene segments in each developing T cell, rather than by expression of germ-line encoded receptors, meaning that even two genetically identical individuals will possess different TCR repertoires.

The TCR is a heterodimer, and in most T cells is made of an α chain and a β chain. Each chain is created through a series of somatic recombination events while the T cell is developing in the thymus, which rearrange one of a number of variable ('V') and joining ('J') gene segments (and in the β chain only, one of a number of diversity ('D') segments) together while the intermediary DNA encoding the other gene segments is excised [8]. This rearranged DNA is transcribed and spliced to a constant ('C') gene to give the receptor chain (Figure 1.2). The gene recombination is performed by enzymes, particularly RAG, recognising recombination signal sequences ('RSS') that flank each of the V, D and J genes.

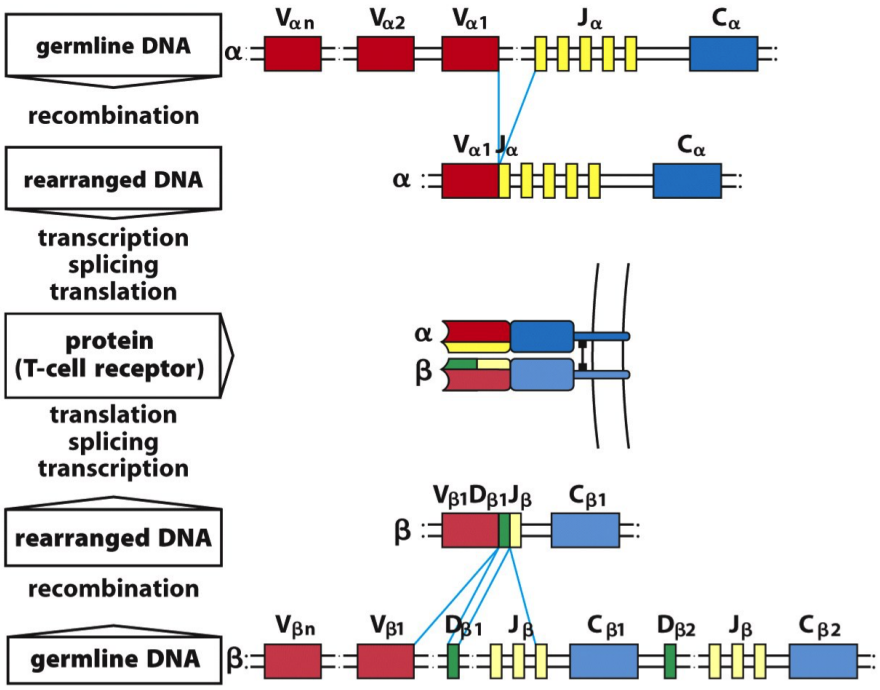

Figure 1.2: Recombination of TCR gene segments
Taken from [107]
Schematic of TCR α and β chain rearrangement.

In addition to the combinatorial potential to create a large number of different TCRs from the number of different V, D and J genes available in the recombination process, additional receptor diversity is introduced by the imprecise nature of the recombination machinery. At each of the junctions between gene segments (V-J for α, V-D and D-J for β), the enzymes involved in the recombination process delete a varying number of the germline nucleotides and include a varying number of non-templated nucleotides [48], allowing for a massive number of potential TCR types.

One study has estimated that there are 10^{15} potential different $\alpha\beta$ TCRs [37], using a combinatorial approach considering the gene segment options plus the effect of nucleotide deletions and non-templated nucleotide insertions, while Janeway's Immunobiology [107] calculates a theoretical $\alpha\beta$ diversity of 10^{18}. Both of these estimates of the number of potential TCRs vastly exceed the number of T cells in a human so this level of diversity will never be realised in an individual. An early TCR repertoire study [11] estimated that there are approximately 10^6 distinct β chains in human blood, each of which paired with on average 25 different α chains, giving a lower bound on the diversity that is realised in a single individual as 2.5×10^7, demonstrating

the vast difference between the theoretical and realised diversity of the TCR repertoire. More recently, a mathematical model has been used to estimate the number of distinct clonotypes present in a human [90], finding that the mean clone size, over the life of a clonotype, is of the order 10. Therefore, since a human contains in the order 10^{10} T cells an estimate of 10^9 distinct clonotypes is reached.

Given the difference between the potential and realised TCR diversity, only a small fraction of the possible TCRs will be rearranged in any individual. With these rearranged TCRs, the immune system needs to be able to provide broad coverage of antigens that indicate the presence of dangerous pathogens that the individual might encounter, while avoiding a harmful immune response to self-peptides that are continuously presented on MHC by APCs. To achieve this, the developing T cells pass through two rounds of TCR-dependent selection in the thymus, as reviewed in e.g. [106].

First, developing T cells encounter thymic epithelial cells that present a selection of self-peptides loaded into MHC ('spMHC'), and they must bind with strong enough affinity to receive necessary survival signals to pass positive selection [80]. Secondly, if a T cell binds too strongly to spMHC complexes presented by thymic APC it is deleted under negative selection [72]. To create tolerance to the full range of self-peptides that peripheral T cells might encounter, the AIRE transcription factor causes expression of a full range of tissue-specific proteins within thymic epithelial cells [9]. Together, these selection processes ensure that the cells that reach the periphery are enriched for those that are able to recognise peptide loaded into MHC molecules without recognising too strongly 'harmless' self-peptides.

However, it is now recognised that the thymic selection process is somewhat 'leaky' [163] and there are self-reactive T cells present in the periphery, some of which may be at frequencies equivalent to those of nonself-reactive T cells. As such, there needs to be mechanisms in the periphery to ensure the T cell population does not cause autoimmunity through recognition of and response to self-antigens, discussed further in Section 1.1.4. In Chapter 3 we develop a model of maintenance of peripheral immune tolerance and investigate the effects of this model on the repertoire.

1.1.3 Clonal differentiation and expansion

When a T cell encounters cognate antigen (pMHC which is 'recognised' by its TCR with enough binding strength), in the presence of costimulation (e.g. from CD80 and CD86 on the antigen presenting cells) and additional 'danger' signals [96] from the innate immune system acting either on the APC or the lymphocytes directly, it becomes activated. Activation of T cells initiates signalling cascades which result in cell differentiation and clonal expansion [27]. Some cells in the activated clone become long-lived memory cells, allowing a more rapid immune response if the same pathogen is encountered by the individual again. Others become effector T cells, providing a targeted, antigen-specific immune response against the pathogen that activated the clone.

A key feature of this clonal expansion is that all progeny of a given T cell carry the same TCR, allowing them to effectively act to clear, in a targeted manner, the pathogen that initiated the immune response. This clonal expansion also means that the repertoire of TCRs (the size and type of each T cell clone) in an individual contains information about the current and historical immune challenges the individual has faced, and may be able to predict the ability of their immune system to effectively counter future immune challenges. A combination of experimental and computational techniques is necessary to extract and interpret such information.

1.1.4 Peripheral regulation of the T cell population

Thymic selection ensures that peripheral T cells are enriched for those bearing TCRs which are able to recognise, but which do not respond too strongly to, self-peptides presented in the context of MHC. However, this system is not perfect, and as such both T cells that are self-responsive and T cells that are unable to recognise MHC-presented peptides may be present in the periphery. In order to maintain an optimal T cell population and to avoid harmful auto-immunity, both positive and negative processes of T cell population regulation are required outside the thymus.

In [142], thymuses from wild type mice are grafted into RAG deficient mice, who do not produce their own mature T or B cells, that either do or do not express MHC class II molecules. It was found that newly generated $CD4^+$ T cells are able to proliferate

in the periphery even in the absence of MHC class II, but that long term maintenance of the peripheral pool of $CD4^+$ T cells is compromised in those mice. In [112] it is demonstrated that the presence of MHC Class I expressing cells is required for $CD8^+$ T cells to successfully survive in the periphery. Both these studies suggest a form of positive regulation of the peripheral T cell population, with T cells that are not able to interact with self-peptides presented in the context of MHC molecules being deleted.

More recent evidence of the requirement for naive T cells to recognise self-peptide MHC complexes is reviewed in [141], including the observation that maintenance of the naive T cell population in the absence of MHC appears to be very context dependent. For example, the studies mentioned above used lymphopenic mice, in which the response of T cells to signals that normally induce homeostatic proliferation is known to be altered. In a study where polyclonal naive T cells are transferred into various non-lymphopenic mice, it is seen that in MHC-deficient hosts phosphorylation of TCRζ (as a measure of TCR signalling) in donor cells is rapidly lost but survival over a month is equivalent to the survival seen in wild type hosts [41], suggesting that interactions between TCR and MHC are not essential for survival of naive T cells.

In [23], a review of experimental evidence demonstrates considerable support for the idea that survival of naive T cells requires survival signals from MHC in conjuction with IL-7 signalling, and that there is considerable competition in the periphery for self-peptides presented on MHC. However, maintenance of the memory T cell population does not require MHC signalling.

If recognition of self-peptide-MHC complexes is required for long-term survival of naive T cells, the homeostatic abundance of a T cell clone in the periphery should be dependent on availability of peptide that its receptor is able to interact with. It has been shown experimentally that the precursor number does affect the maintenance of naive and memory T cells in non-lymphopenic hosts [57]. Cells of clones with lower abundance are more likely to survive than cells from large clones, suggestive of intra-clonal competition for peptide and therefore of dependence on MHC for T cell maintenance. Competition between different clones for peptide has been modelled [138, 137], showing that clones which have greater overlap of self-peptides recognised with other

clones tend to become extinct more rapidly. This competition for peptide results in a diverse repertoire, where the most successful clones occupy 'niches', with few other clones recognising the same self-peptides. An alternative modelling approach is used in [97], where the 'optimal' repertoire is found to be sparse when cross-reactivity between clonotypes is present.

In addition to positive regulation of the peripheral T cell population, removing cells which are unable to recognise peptide in the context of MHC, negative regulation is also required in order to avoid dangerous immune response to self-peptides. Avoidance of self-response is known to involve regulatory T cells ('Tregs'), which suppress the function of conventional T cells. The role of Tregs in the maintenance of self-tolerance and the avoidance of autoimmunity has been thoroughly reviewed (e.g. [128, 129]) and is not considered further here.

The 'danger model', first proposed by Matzinger in 1994 [95], suggests that APCs are activated by alarm signals from injured or infected cells, and that only TCR interactions with peptide presented by activated APCs are able to stimulate a T cell response. A possible extension of this model is that APCs that have not been activated by these alarm signals would have the potential to be tolerogenic, ensuring no immune response is initated following interaction with the peptides being presented on resting APCs. The evidence for the existence of tolerogenic DCs is conflicting, but as reviewed in [135] they were first suggested in mouse experiments where an antigen was delivered to DCs without the DC becoming activated [58]. This resulted in proliferation of antigen specific T cells, rapidly followed by contraction of the population and the remaining cells becoming unresponsive to later antigen challenge. However, the evidence for true tolerogenic DCs is still relatively unclear. An additional role of DCs in peripheral regulation of the T cell population is the ability of immature DCs to induce Tregs, as discussed in detail in [92]. Dendritic cells have also been shown to be able to delete $CD8^+$ T cells [166] directly through an MHC dependent mechanism, and a particular subpopulation of DCs which express perforin, a cytolytic protein that forms pores in the cell wall of target cells, have been shown to be important for avoidance of tissue-specific autoimmunity [167].

In addition to dendritic cells, other APCs are able to regulate the peripheral T cell population. In secondary lymphoid organs resident stromal cells are involved in recruiting and organising T cells, so that they are able to meet DCs presenting antigen but additionally, as reviewed in [149], they express and present peripheral tissue restricted antigens, and in doing so tolerise naive T cells to these peptides. There is also considerable debate about the role of Langerhans cells (specialised DCs present in the skin and mucosa) in T cell tolerance, with conflicting evidence regarding the role of these cells in the epidermis and their mechanism of action. In some experimental models Langerhans cells have been found to stimulate T cell responses while in other settings they are seen to induce tolerance, as reviewed in [66, 89].

Despite understanding and modelling some of the mechanisms of positive and negative regulation of the T cell population in the periphery, understanding how the immune system is able to make a robust decision about whether to initiate a response to a presented peptide remains a challenge. The flexibility of the TCR-pMHC interaction results in much cross-reactivity, and there is overlap in the distribution of affinities with which TCRs bind to self or nonself presented peptides. One approach to understanding this is to consider cooperative behaviour between T cells and dendritic cells, and to consider the decision making process as occurring not at a single cell level but at a population level. In [28] an elegant model of quorum-sensing is described, where an immune response is only initiated if sufficient T cells recognise presented peptide on an APC. Once positive and negative thymic selection has skewed the distributions, across the whole T cell population, of affinity to self and nonself peptides, a carefully selected quorum threshold is then able to accurately discriminate between presented peptides in the periphery.

In Chapter 3, concepts of tolergenic APCs, cooperation between cells of the immune system and cross-reactivity between T cell clonotypes are combined in a model of maintenance of immune tolerance at steady state in the periphery.

1.1.5 Studying the TCR repertoire

The majority of studies into the role of T cells in an immune response either consider the T cell population without considering clonal distinctions (e.g. measuring population subset sizes) or investigate the role of antigen-specific lymphocytes in isolation, either in vitro (e.g. using tetramers to isolate antigen-specific T cells) or in vivo using TCR transgenic approaches.

In contrast, deep-sequencing studies covering the whole TCR repertoire provide a holistic approach to investigating the role of the T cell population at steady state or in response to an immune challenge, allowing systems-level questions to be asked about how clones interact and whether these interactions affect the behaviour of the whole system.

1.2 TCR repertoire sequencing and analysis pipeline

Initial studies into the TCR repertoire provided only low-throughput and/or coarse data regarding the receptors present in a sample. For example, spectratyping techniques using Vβ specific primers allowed the distribution of lengths of CDR3 within a Vβ family to be measured [51, 52] but did not provide any sequence information regarding the different clonotypes within the V gene family. Before the development of high-throughput sequencing, to obtain sequence information about individual TCR clones individual antigen specific T cells were isolated using tetramers. These single cells were then grown up into T cell clones and their receptors sequenced via Sanger sequencing.

The development of high-throughput sequencing ('HTS') has provided the ability to acquire the nucleotide sequence of millions of molecules in one experiment, and protocols have been developed to apply this technique to libraries of TCRs. These sequencing data provide a more comprehensive and detailed view of the TCR repertoire in a sample than was previously possible, but require careful processing and analysis to obtain biologically relevant information.

1.2.1 Library preparation protocols

Various TCR sequencing protocols have been developed, but all need to include steps to extract the template nucleic acid material from the sample and amplify the template molecules via PCR to obtain concentrations suitable for loading onto the sequencing machine.

TCR sequencing protocols either use genomic DNA (gDNA) or messenger RNA (mRNA) from cells as the template to be amplified and sequenced. The choice of starting material affects the experimental protocol, due to different priming sites available for PCR amplification, as well as data processing decisions, discussed later.

TCR gDNA includes intronic regions, which are removed by splicing following transcription and so are not present in mRNA. The most substantial of these introns, and the most relevant in the development of TCR sequencing protocols, is the section between the J gene and the constant region which can be long, especially in comparison to the capabilities of the PCR reaction and of sequencing machines. The presence of this intron means that amplification from gDNA is not feasible using primers against the constant region, shared by all template molecules, and still ensure coverage of the CDR3 which is generally the region of interest. Protocols sequencing from gDNA [125] therefore generally use a panel of J region primers, covering the known functional J genes, to avoid amplifying the intron region.

Similarly, amplification of TCR target molecules, whether from mRNA or gDNA, is complicated by the lack of a conserved sequence at the V-gene end of the molecule that could be used as a primer target. This can be solved by using a panel of V primers covering the known functional V genes. As an alternative to using a panel of V gene primers, an adapter containing a known sequence can be added to the V end of the molecule to be used as a priming site. The addition of the adapter is usually performed through a technique known as template switching (e.g. [45, 156] and others), which exploits a property of some reverse transcriptase enzymes, allowing the additon of non-templated nucleotides to the end of the molecule, providing a priming site. Alternatively, as in the protocol developed in our lab, a ligation step can be used to add a primer site onto the V end of the template molecule. Once a known sequence has been added to each tem-

plate molecule, PCR amplification for mRNA protocols can be performed between this primer site and a primer site in the constant region of the TCR, meaning only one pair of primers per constant region (α or β) is needed to amplify all TCR rearrangements found in the sample.

In our lab, the TCR sequencing protocol, as described in [59], uses mRNA and a ligation approach to add a known primer site to the V end of the target molecules and amplification of molecules using primers between this site and the constant region. The mouse TCR repertoire data analysed in Chapter 4 from Nir Friedman's group at the Weizmann Institue is sequenced from mRNA using a PCR reaction between a panel of V primers and the constant region as described in [91].

One of the difficulties with the current high throughput TCR sequencing protocols is that they allow data to be collected regarding both the α and β TCR chains that are present in a sample of cells, but do not allow identification of which α was paired with which β in a cell. Protocols obtaining paired chain sequencing data have been developed, which either physically link the α and β RNA from a cell before sequencing (e.g. [148, 99]), amplify both the α and β chains from single cells and Sanger sequence (e.g. [36]) or alternatively apply a unique cellular barcode to RNA from each cell before sequencing (e.g. [55]). However these techniques are currently much lower throughput and more expensive than the protocols obtaining unpaired sequencing data.

Alternatively, computational approaches are being utilised to obtain paired chain data. PairSEQ [63] is a method that relies on the almost unique nature of clonal TCR rearrangements within a sample. The cells in a sample are split into multiple subsets and α and β chains present in each subset are sequenced following a standard TCR sequencing protocol. The α and β chains that were present together in a cell should then appear in the sequence data from the same subsets, and from no other subsets. In this manner a combinatoral approach can be used to identify the $\alpha\beta$ pairings present in the sample, at lower cost and with higher throughput that the paired sequencing approaches.

1.2.2 Processing of HTS TCR data

After TCR molecules have been sequenced, the sequence data needs to be processed into a usable form before further analysis can be performed. The data processing can be thought of as having three steps: (i) identification of the TCR clone (either at the rearrangement-event level or the CDR3 level) in each sequence read, (ii) sequence error correction and (iii) quantification of clone sizes.

1.2.2.1 Identifying clonotypes from sequence data

In order to perform analyses of the TCR repertoire from HTS data, first the receptor rearrangements, or the CDR3 sequences, corresponding to each sequence read need to be identified. For applications of HTS to germline encoded products, the identification of the relevant biological product in sequence data is a relatively straightforward sequence alignment problem, only made difficult by the quantity of sequence reads involved. Techniques such as NCBI's BLAST and its extensions allow for rapid identification of the portion of the genome that a sequence read covers.

In order to identify the TCR rearrangement present in a sequence read, the read needs to be matched to one of a number of V genes, which have very similar nucleotide strings, and to one of a number of J genes which also have high similarity. Additionally, the sequence data from the TCR rearrangement will contain an unknown number of non-templated nucleotides for which no alignment to a genome is possible. Both of these features make it more difficult to identify the appropriate TCR rearrangement for each sequence read, and more diffcult to correct sequence error.

The common approach to this problem is to determine the rearrangement events that a sequence read represents by first identifying the constituent V and J genes. Once these have been decided then the number of nucleotide deletions and non-germline insertions can be predicted and the nucleotide or amino acid sequence of the CDR3 determined. Methods for identifying the genes present in a TCR sequence read are discussed below.

It should be noted that the identification of the rearrangement that occurred in the developing thymocyte to produce a particular TCR is not a trivial problem because it is

apparent that there is not a one-to-one relationship between rearrangement event and sequenced TCR: the same nucleotide sequence can be created using different V or J gene segments, different numbers of nucleotide deletions and different non-templated nucleotide additions. This is referred to as 'convergent recombination' and in [42] probabilistic models are used to infer likely rearrangement events from observed nucleotide sequences. If the sequenced TCR data is being used to infer properties of the rearrangment mechanism in developing thymocytes the probability of each of the possible rearrangement events is important to consider. However, if the characteristics of the existing TCR CDR3 repertoire are being investigated it may be sufficient to assign each sequence read to just one possible rearrangement event.

In order to determine the V and J genes present in the sequence read, some form of string-matching algorithm is needed. Many pipelines make use of pairwise local alignment algorithms to obtain a score for each possible gene and assign the read to the highest scoring alignment (e.g. [85, 155]). IMGT/HighV-QUEST [5] is a high-throughput implementation of IMGT/V-QUEST [26] in which assignment of gene segments to sequence reads is performed by pairwise alignment techniques. It can be accessed through a web-based interface, and is used in a number of studies including [86, 151].

Alternatives to pairwise alignment are also used to identify V and J genes within a TCR sequence read. MiTCR [21] is software that assigns gene segments to high-throughput TCR sequencing data by first searching for pre-specified 'seed' n-mers (generally covering the ends of the CDR3 region from each gene segment) in the sequence read and then attempting to extend the alignments to identify the genes that are present, with each potential gene segment being given an aligment score. In our lab, Decombinator [145] is used to process HTS TCR data. Decombinator utilises the Aho-Corasick algorithm [3] which constructs a finite state machine to enable efficient searching of all sequence reads to look for instances of a set of 'tags' which uniquely define the possible V and J gene segments.

1.2.2.2 Sequence error correction

The presence of erroneous base pair calls in sequence data due to PCR or sequencer error is a substantial issue, with up to 6% error being observed [113]. Different approaches can be taken to try to minimise the impact of this sequence error on repertoire analysis. Common approaches are outlined below.

Sequence reads in fastq format come with an attached quality score for each basepair, indicating the probablity of a correct basepair call. Many groups apply a filter to sequence reads depending on this quality score, either requiring a minimum average quality over all base pairs or a minimum per-base quality required. It is also possible to incorporate the quality scores into the assignment of V/J genes, allowing for mismatches where quality score is low [155]. In addition, many analysis pipelines choose to discard any read where there is an uncalled basepair, any read which is out-of-frame, or any read which does not contain the C-region primer or the expected multiplexing index (e.g. [79, 151]).

After assignment of V/J genes and identification of the CDR3 contained in each read, some groups restrict which TCR clones are taken forward for further analysis by discarding 'small' clones, considered to be prone to being counted as a distinct clone due to sequence error rather than a true biological difference. Those clones which are discarded as being 'small' can be defined in a number of ways. Often, clones that appear below a certain number of times are discarded automatically, and in [164] those clones which appear below $0.5\times$ 'coverage' are abandoned, where coverage is defined as the average number of annotated sequence reads per cell of input material. Alternatively, in [156], only those largest clones which account for 96% of the sequence reads are retained, where 96% has been determined to be the 'best' cutoff to remove error by analysis of the J genes in each sequence.

Instead of discarding sequences containing error, and therefore potentially losing quantification information regarding the clones, another approach commonly used is to group the reads from the same clones and then incorporate satellite reads into larger clones (or discard the smaller clones) according to some algorithm. [124] uses a nearest neighbour algorithm to group reads together into a single clone, while [164] discards

small groups of reads if they are different from a larger group in just one position in the CDR3 and if the number of reads in the smaller clone is $< 5\%$ of the number of reads in the larger clone.

In [22] an error correction algorithm is proposed that starts with creating groups of 'core' clonotypes, containing reads with high quality sequences and identical CDR3s. To correct for errors from the sequencing machine, low quality reads are mapped to these core clonotypes with mismatches allowed in positions where the sequence quality is low. Then, to correct for errors introduced during PCR, sequence reads are grouped into larger clonotypes if the only mismatches are in the V or J regions. This approach has been packaged into a piece of software, MiTCR [21].

1.2.2.3 Quantification of clone sizes

Many analyses of TCR sequence data are interested in the size of different clones, both relative to other clones in the same sample or relative to the same clone in a different sample either from a different individual or at a different time.

Simply counting the number of times each TCR clone or each CDR3 appears in a sample, after correction for sequence error using one of the methods described above, can give an estimate of clone size in the sample, but it is important to consider the effect of the starting material choice on the interpretation of this number. Sequencing from gDNA ensures that one copy of the rearranged TCR is observed per cell in the sample, whereas there are multiple copies of mRNA per cell. Observations from our group have suggested that the number of TCR mRNA molecules per cell are in the range 1 - 10 and it has been suggested that the per cell TCR mRNA count may vary with antigenic stimulation [114]. If we can assume that mRNA levels are relatively stable within a cell and consistent between cells in a sample then counts of sequenced mRNA molecules are a proxy for clone size in the same way that sequenced gDNA molecules are, otherwise counts of mRNA molecules should be interpreted as a function of clone size and receptor expression levels.

Protocols that use panels of PCR primers against the V and/or J genes need to account for the potential primer bias that these could introduce. In [111] the bias resulting from

PCR using a panel of V region primers is corrected for via a probabilistic method of normalising clone size in sequence data based on sequencing of a synthetic libary of TCRs. In [30], synthetic templates are used to measure the bias in PCR using panels of primers against both the V and J regions. It is found that each gene region has a biased amplification but that there is no interaction between the two gene segements, and an optimal primer mix, minimising the PCR primer biases, is found through titration experiments. The residual bias using this optimised primer mix is measured using the synthetic library and this gives a set of normalisation factors which are then applied to sequenced TCR repertoire data.

Quantification of clone size is also complicated by the inherent stochasticity of the PCR amplification process. Efficiency of amplification by PCR depends on a number of factors, including the GC content of the template. In a heterogenous mix of template, such as a sample of the TCR repertoire, it can't be assumed that each initial template molecule will amplify at the same rate. Additionally, the stochasticity of the PCR produces a certain amount of noise in the amplified sample. In an experiment where the same sample of T cells in blood is split into two and each is amplified under the same conditions [124] the observed clone sizes in the two split samples are seen to be relatively uncorrelated for 'small' clones (observed < 100 copies per sample).

Given this observed heterogeneity, the amplified pool of molecules that are sequenced may not be quantitatively representative of the original sample which might affect conclusions drawn from analysis of the data. In order to account for this heterogeneity, some antigen receptor sequencing studies implement a unique molecular barcoding approach where each molecule in the initial sample is uniquely tagged with a nucleotide string. The particular techniques used to label molecules are discussed in Chapter 2. The unique barcodes allows reads from the final sequencer output to be clustered according to which initial molecule they are derived from, meaning that variable PCR amplification efficiency can be corrected for and a more accurate estimate of initial clone size can be obtained.

In Chapter 2 we investigate the bias that PCR amplification heterogenity might introduce into TCR repertoire sequence data and suggest that single molecule barcoding is

vital to ensure quantitative analyses are robust.

1.3 Analysis of TCR repertoire sequence data

Many studies inferring properties of the immune system from TCR repertoire sequencing data can be thought of either as studies into the system-level properties of the T cell population at steady state or under immune challenge, or as studies attempting to identify characteristics of an antigen specific response in the repertoire. We discuss both of these approaches to the study of the TCR repertoire below.

1.3.1 Global properties of the TCR repertoire

Global properties of the TCR repertoire (including diversity, gene segment usage distributions, and repertoire similarity between different samples) are often analysed as indicators of the health, effectiveness or immunocompetence of the T cell population. Many studies consider the TCR repertoire of healthy individuals, to gain insight into the mechanics of the formation of the repertoire or the baseline state of the repertoire, while others consider the repertoire in various disease settings.

1.3.1.1 Diversity: richness and evenness

Many early TCR repertoire studies focussed on estimating the diversity of the repertoire in a healthy individual and additional studies have considered the effect of immune challenge on the diversity of the repertoire, to measure how disease disrupts the T cell population.

Diversity of the repertoire can be thought of in a number of ways, and can be quantified using a number of measures. Two facets of diversity are ‘richness’ and ‘evenness’, and most measures of diversity combine both of these. The richness of a repertoire describes how many clones are present in a sample. This can most simply be measured by counting the observed clones and comparing this number between samples. However this number will depend both on sample size, since the number of small clones that are observed will be affected by sampling, and on sequencing depth, since not all

clones present may be sequenced. The evenness of a repertoire describes how uniform the abundance of the different clones are within the population.

The estimated richness of the TCR repertoire in a healthy individual is discussed in Section 1.1.2, and in [79] it is found that the memory compartment is approximately half as diverse (in terms of TCR richness) as the naive compartment in $CD4^+$ T cells, and between 1/3 and 1/10 as rich in $CD8^+$ T cells.

The evenness of a TCR repertoire is measured in many studies using the Gini index, an equality measure commonly used in economics and social sciences which describes how equally a resource is distributed amongst a population. A Gini index of zero corresponds to a completely equal distribution of resources and a value of one corresponds to a maximally unequal distribution. In the case of TCR repertoires, it can quantify whether a population is dominated by a few large clones or if most of the clonotypes are of equal size.

A number of diversity metrics, often borrowed from other disciplines, are used in this work and in other studies to describe and compare the diversity (combining richness and evenness) of sequenced TCR repertoires. The Shannon Index, or Shannon Entropy, was first developed with reference to the information in strings of text. In the context of TCR repertoires, it can be thought of as describing the uncertainty involved in predicting what clone a randomly selected read will belong to and is calculated as:

$$H' = -\sum_{i=1}^{R} p_i \ln p_i$$

where R is the number of different clones observed and p_i is the proportional abundance of the ith clone. Sometimes the Shannon Entropy is normalised by dividing by the log of the number of observed clones. Another metric of diversity is the Simpson Index, which calculates the probability that two TCRs, selected at random from a dataset, are of the same clone:

$$\lambda = \sum_{i=1}^{R} p_i^2$$

In a highly diverse sample the Simpson Index gives very small values so often the inverse Simpson, $1/\lambda$ is quoted.

1.3.1.2 Recombination biases

In [102], many examples of pre-HTS studies demonstrating biases in the recombination of TCRs are reviewed, including studies showing that different promoter sequences upstream of V genes affect the rates at which each is expressed and showing that there is preferential pairing between V and D/J gene segments. All of the biases in the recombination machinery result in non-uniform probabilities for each of the potential TCRs to be rearranged, reducing the diversity of the repertoire.

In [108] a comprehensive study of the recombination events in non-productive rearrangements (to avoid effects of selection) from HTS data of the TCRβ repertoire is undertaken to understand biases in the rearrangement of TCRs. The probabilitistic properties of the recombination process are inferred, using a maximum likelihood approach since single multiple recombination events can lead to the same nucleotide sequence. It is found that the frequencies of V and J genes vary significantly (as many other studies have found and in [111] is attributed to the physical conformation of chromatin), but interestingly no dependence between V choice and J choice was observed, suggesting that previously reported correlations might be a function of selection pressure rather than the recombination process. It is found that the number of deletions from the V and J genes is dependent on the gene used, while the VD and DJ insertions are uncorrelated to each other, and the probability of a particular nucleotide insertion depends strongly on the 5' nucleotide. The derived probabilities are remarkably consistent between individuals, with the greatest variability seen in the gene usage probabilities. This consistent probabilistic model of recombination both reduces the potential repertoire diversity and is able to account for the number of sequences shared between two individuals.

1.3.1.3 Repertoire similarity and public clonotypes

Given the massive amount of potential diversity in the TCR repertoire, it might be expected that no identical TCRs are shared between two individuals. However, the biased nature of the recombination machinery, as well as the requirements for a TCR to be functional and able to recognise peptide fragments presented in MHC molecules, im-

pose limits on diversity and suggest that the same TCR may be generated and selected in multiple individuals. Clonotypes shared between many individuals are described as 'public', and have been consistently observed in HTS TCR repertoire data. The extent that public clones arise simply due to highly probable recombination events or whether convergent selection (either thymic or post thymic) is required to explain their presence remains unclear.

The overlap between the sets of clones observed in two samples is often calculated as a metric of repertoire similarity between two individuals or between two samples from the same individual. The Jaccard Index is defined as the size of the intersection between two sets divided by the size of the union i.e. $J(x,y) = \frac{|\cap(x,y)|}{|\cup(x,y)|}$. In terms of TCR repertoire, it calculates the proportion of distinct clones (total observed across either sample) that are observed in both samples. In some studies an extension to this similarity metric is used, where the abundances of the clones each of the samples are used to weight the Jaccard to give more importance to highly abundant clones. Alternatively, the Battacharyya coefficient is another measure of the similarity of two sets which takes into account the clonal abundances, and is calculated as $BC(x,y) = \sum_i \sqrt{x_i y_i}$ over all clones i present in either sample, where x_i, y_i are the abundances of clone i in the respective samples.

In [126], the naive $CD8^+$ subsets of TCRβ sequences are compared for every pair of seven individuals, finding at least 100,000 CDR3βs shared between each pair, including a pair with no shared HLA alleles. The sharing within the memory compartment was smaller, although there were still at least 1000 shared CDR3βs found between each pair. The number of shared sequences in the naive compartments for a pair of individuals agrees well to the prediction made via a model of biased VDJ recombination. In the same vein, [156] finds that in a pair of HLA-matched individuals, approximately 14% of the CDR3β repertoire is shared. Shared TCRβs were found to tend to have a shorter CDR3 length as well as being more abundant than private sequences.

In another study considering the overlap in TCR repertoires, the naive and memory compartments in four individuals [153] were sequenced. In this study, a lower amount of sharing was observed, with less than 10% of the unique TCRβ amino acid sequences

from one individual being shared with another individual. The abundance of the clone was again found to be correlated to sharing, both between subsets and between individuals, showing that public clonotypes are more abundant. In [42], models of V(D)J recombination suggest that observed public sequences are due to chance during the biased recombination process rather than any convergent selection. It has been suggested that public clonotypes might be involved in autoimmune responses [91] but the function of these T cells is still not fully understood.

In a study of antigen-specific T cells in mice, analysis of tetramer-sorted influenza $CD8^+$ T cells [152] from C57BL/6J mice also showed that highly shared identical sequences (with identity being defined as Vβ, Jβ and CDR3β amino acid sequence matches) have fewer nucleotide additions, and highly shared sequences are present with more different encoding nucleotide sequences, suggesting a high level of convergent recombination. A similar finding is demonstrated in full repertoire analysis (rather than just within a tetramer positive subset) in [91].

1.3.1.4 Perturbations to the TCR repertoire in disease

Diversity and disease

The diversity of the response to a particular pathogen is thought to be related to disease outcome. C57BL/6 mice infected with a herpesvirus (HVH-1) have a poorer survival rate than mice (referred to as 'bm8' mice) that differ in the MHC class I molecule by 4 amino acids within the peptide binding site [100]. This difference is found to be $CD8^+$ T cell dependent, and the T cell response of C57BL/6 mice mostly utilises TCRs of the Vβ10 and Vβ8 families, while the bm8 mice use a broader range of Vβ families, and more diverse clonotypes within the families.

In [88], spectratyping of the TCR repertoire is used to show that TB patients have a more restricted repertoire than healthy volunteers, and the complexity of a patient's repertoire is negatively correlated with their disease severity, suggesting that a diverse TCR repertoire may be required for control of disease.

Diversity as a measure of success of treatment

In a study of the TCR repertoire in peripheral blood of patients undergoing anti-CTLA4 treatment (tremelimumab) for cancer it was shown that there was a significant increase in richness (measured by the absolute number of TCR rearrangements observed) after treatment for the majority of patients [122]. However, there did not appear to be a change in the distribution of the repertoire, as measured by Pielou's evenness index. The increase in richness was not correlated to the response of the patient to treatment, but those patients with most increased richness were those that experienced most toxicity, leading to the suggestion that the increase in richness of the repertoire might be non-specific expansions, perhaps related to autoimmune clones, rather than expansion of clones against the cancer.

The TCR response to anti-PD1 treatment (pembrolizumab) for metastatic melanoma is considered in [147]. Regression of tumours is found to be associated with proliferating $CD8^+$ T cells localising to the tumour, and when the β chains of the TCRs present in the tumour pre-treatment are sequenced it's found that a higher 'clonality' is indicative of a better response on treatment. In this study, the clonality metric is calculated as $1-$ normalised entropy, where normalised entropy is the Shannon entropy divided by the log of the number of different clones observed. In the same study, comparison of the TCRβ clone sizes pre- and during treatment shows that the patients that successfully respond to the anti-PD1 treatment have $10\times$ more clones expanding during therapy.

A HTS TCR study of patient samples following haematopoietic stem cell transplantation ('HSCT') [151] showed that TCR repertoire diversity, as measured by the inverse Shannon, was reduced in comparison to healthy volunteers following transplant with a T cell depleted graft and even a year after treatment the population diversity had not recovered. However, this appears to be dependent on whether $CD4^+$ or $CD8^+$ cells are considered. The T cell repertoires in patients receiving T-cell depleted autologous stem cell transplants to treat systemic juvenile idiopathic arthritis are studied in [161]. Spectratyping showed that the diversity in the $CD4^+$ TCRβ repertoire is restored by 12 months after transplant, while diversity is only restored in some of the $CD8^+$ Vβ families by the same time point. High throughput sequencing of samples indicates that some clones appear to survive the conditioning process, or are grafted back into the patient during transplant.

In work from our lab [59], the TCR repertoire in HIV+ patients before and during antiretroviral treatment ('ART') was studied, showing that HIV+ patients have a perturbed repertoire, with much lower diversity than healthy controls and that this is not restored after three months of ART despite viral loads becoming undetectable.

TCR repertoire sequencing providing insights into disease mechanism

In a study of TCR repertoire in colorectal tumours [131] it is seen that although the absolute diversity (the number of distinct TCR clones observed) is similar between tumour and nearby mucosal tissue from the same patient, the normalised Shannon is higher in the tumour. This suggests that the repertoire present in the tumour is less polyclonal than that found in surrounding tissue, perhaps due to antigen specific T cell recruitment to the tumour or retention in the tumour. However, the amount of repertoire overlap between tumour and nearby tissue in this study was not related to the physical distance between the two samples. Additionally no differences in V and J gene usage or CDR3 length were found between tumour and nearby tissue repertoires, suggesting that the observed difference in normalised Shannon cannot be fully explained by a skewed repertoire at the global level.

In a study considering the repertoire in patients with rheumatoid arthritis, samples were taken from multiple affected joints as well as from blood [78]. The highly expanded clones are shared between different affected joints in a patient, but are not shared between joints and blood, again suggesting antigen specific recruitment or retention.

In a study of the repertoire in ovarian cancer [44], multiple samples from primary tumour, metastatic sites and blood are taken from each patient. A weighted Jaccard approach is taken to measure the similarity between repertoires observed at different sites, and it is found that within a single metastatic site the repertoire is very similar, but between a metastatic site and the primary tumour there is less similarity. All the repertoires from tumour samples show little similarity to the repertoire in blood.

1.3.2 Antigen specific T cells

The global properties of the whole TCR repertoire are able to provide information regarding the status of the immune system, as discussed above, but considering the properties of the individual T cells that comprise the immune response to a particular antigen can provide additional information regarding disease status, correlates of protection and patient stratification. Even when an epitope (a specific peptide/MHC complex) is known, it is not currently known how to map this to the TCRs that will recognise it, and in many disease situations the antigenic peptide is unknown or prone to mutations. As such, experimental techniques to identify, isolate and sequence responsive T cells are used in many studies to answer questions regarding the magnitude, breadth and effectiveness of the T cell response.

Identification and tracking of antigen specific clones in diagnosis and treatment

One of the first applications of HTS of TCRs was in T cell leukaemias, where T cell clonal expansion occurs independent of antigen. In the context of cutaneous T cell lymphoma ('CTCL'), the TCRβ and TCRγ repertoire is sequenced from punch biopsies [76], demonstrating that the normalised entropy (referred to as the 'clonality') of the full repertoire is correlated with disease severity, with the biopsies from more advanced disease having a less polyclonal repertoire. This study also showed that the repertoire in a skin lesion from a patient with CTCL is dominated by the single malignant TCRβ clone. This is also the case in other diseases, such as psoriasis and dermatitis, but is more severe in CTCL and the proportion of the repertoire occupied by the largest clone can distinguish between CTCL and other skin conditions. The malignant clone in CTCL patients could be tracked over time and in multiple sites in the same patient, informing clinical treatment, for instance in one patient recurrence of disease after stem cell transplant was identified rapidly by sequencing of the same malignant clone in a skin lesion, allowing prompt treatment. Similarly, identification of a single clonally expanded TCRγ in patients with mycosis fungoides was found to be a more sensitive diagnostic than the current standard, suggesting this could be used to inform earlier or more accurate clinical diagnosis [140].

TCR sequencing data is used in a similar manner to detect minimal residual disease

('MRD') in patients with acute T lymphoblastic leukemia ('T-ALL') [160]. Here, pre-treatment HTS TCR samples were used to identify the TCR sequence of the patients' neoplastic T lymphoblast clones, and HTS of TCR repertoire was then able to detect MRD in post-transplant samples with more sensitivity than the commonly used flow cytometric method. A more recent study [87] has sequenced both TCR and Ig to quantify MRD in blood samples from B-ALL patients undergoing allo-HCT treatment. The quantified level of MRD in pre-conditioning samples is predictive of transplantation failure, while MRD quantified above a particular threshold in a post-transplant sample was 100% positively predictive of eventual relapse. However, patients with MRD quantified below the threshold did not all avoid relapse. These studies suggest that TCR and Ig HTS data could inform clinical intervention before symptoms of relapse present clinically and therefore improve patient outcome.

Graft vs host disease ('GvHD') is a frequent complication after bone marrow transplation, and is the result of inappropriate immune response against unknown self antigens. The standard approach is to treat with steroids, but some patients are not responsive. The TCR repertoire in the blood of GvHD patients, after HSCT, does not predict patients which respond to steroid treatment [101]. Instead, a method tracking particular clones in patients over a timecourse was developed. Patient specific 'indicator clones' were identified as the largest clones in a pre-treatment sample from diseased tissue. Then the number of indicator clones in blood after 30 days of steroid treatment is related to the success of treatment, with the patients who require secondary treatment being those that have fewest present.

Properties of antigen-specific clones

In [81], various functional assays of T cell responses to a CMV peptide are used to sort CMV responsive from CMV non-responsive cells. The TCR repertoires of the unsorted cells as well as the CMV responsive and CMV non-responsive subsets are sequenced. For each of the functional assays considered, those clonotypes that are highly enriched in the CMV responsive sample above the unsorted and CMV non-responsive samples are taken to be CMV specific. There is good overlap in the sequences defined as CMV specific in the sequence data from pentamer staining, CD137 and CD107 functional

assays, showing that the same clones are responsive under different functional tests. In addition, using CFSE staining as the functional assay allows identification of CMV specific TCRs that are initially present at low frequency but proliferate strongly in the presence of antigen, which are otherwise lost in the sequence analysis technique presented in the study. It is found that sets of CMV specific TCRs identified using the combination of functional assay and HTS demonstrate sequence-level similarity between themselves and with previously described CMV responsive clones. However, whether there is a sequence-level motif of CMV-specificity that is strong enough to predict whether a cell will respond to antigen is not explored in this study.

The repertoire of tumour infiltrating lymphocytes ('TILs') in melanoma is studied in [53], and the properties of those TILs that are tumour-reactive are explored. The tumour-reactive $CD8^+$ TILs are differentiated from other TILs by PD-1 expression, and PD-1 positive TILs have a more skewed TCR repertoire, with the repertoire being more dominated by a few large clones, than other TIL subsets. The thirty most frequent PD-1^+ TILs occupy over half of the subset, while the equivalent in the PD-1^- subset is just 5%. Additionally, within the most frequent PD-1^+ TILs are found known mutation-specific clonotypes, while these clonotypes are only found at lower frequency in the PD-1^- subset. This study therefore suggests that PD-1 expressing abundant clones may represent tumour-reactive TILs.

MHC alleles HLA-B*27 and HLA-B*57 have been associated with ability to control viral load of HIV, but most people expressing these molecules are unable to control viral load without treatment. In [33], the differences between T cells from people who express HLA-B*2705 and are either able (controllers) or unable (progressors) to control HIV load without treatment are studied. A dominant epitope of the HIV Gag protein is known to be KK10, but no difference was found in the proportion of $CD8^+$ T cells which are responsive to this epitope between controllers and progressors. Additionally there was no difference seen in functional assays of the responsive cells, suggesting the ability to control HIV viral load is not due to the number of T cells able to respond to an epitope of the virus. However, when cultured with HIV in vivo, T cells from progressors demonstrate much less ability to control viral replication than T cells from controllers, demonstrating that the T cell population does affect the differing

phenotype. KK10 specific T cells from controllers have increased potency and cross-reactivity against naturally arising KK10 mutations when compared to KK10 specific T cell from progressors. Cloning and sequencing of KK10 specific T cells from controllers and progressors was performed, and those clones that were found to be immunodominant in vivo in controllers are found to be most effective at killing HIV infected cells in vitro. This study suggests that the ability of an individual to control HIV infection depends on the functional properties of the particular TCRs of the T cells that are selected for, rather than simply on the total number of clones that are selected.

In a small study of HIV-epitope responsive TCRs in a pair of identical twins simultaneously infected with the same strain of HIV-1 [165] it is found that the T cell clones raised in each individual are entirely private, both against shared and distinct viral epitopes, demonstrating the extraordinary heterogeneity and plasticity of the TCR repertoire, even in genetically identical individuals.

1.4 Scope of this thesis

The aim of this thesis is to explore aspects of how the TCR repertoire is regulated both before and after immune challenge. Specifically, this thesis will:

1. **Address PCR heterogeneity and its effect on quantification of clone size** within high throughput sequenced TCR repertoire data. Amplification of target molecules by PCR is an essential component of most sequencing protocols, but demonstrates variability in its amplification efficiency meaning that the amplified pool is not necessarily quantitatively representative of the original target molecules. In Chapter 2 the extent of this heterogeneity is investigated, and single molecule barcoding proposed as essential to produce reliable quantitative TCR repertoire sequencing data.

2. **Develop a computational model of T cell tolerance** of self at steady state. This model (Chapter 3) incorporates co-operative behaviour between antigen presenting cells and the T cell population, leading to a gradual reshaping of the TCR repertoire in order to avoid excessive total affinity to self-peptides within the

population.

3. **Investigate changes in the TCR repertoire after immunisation** with an adjuvant with or without addition of a model antigen (Chapter 4). TCR repertoires from different mice under different immunisation conditions, and with or without any in vitro culture of cells with antigen, are studied to try to identify features defining the response to antigen.

Chapter 2

Computational analysis of stochastic heterogeneity in PCR amplification efficiency revealed by single molecule barcoding

The work presented in this chapter was published in [18].

The polymerase chain reaction ('PCR') is one of the most widely used techniques in molecular biology. In combination with high throughput sequencing ('HTS'), PCR is widely used to quantify transcript abundance for RNA-seq as well as in the context of analysis of T and B cell receptor repertoires. In this chapter, DNA barcoding is combined with HTS to quantify PCR output from individual target molecules. Computational tools that simulate both the PCR branching process itself, and the subsequent subsampling which typically occurs during HTS sequencing, are developed. We explore the influence of different types of heterogeneity on sequencing output, and compare them to experimental results where the efficiency of amplification is measured by barcodes uniquely identifying each molecule of starting template. This work demonstrates that the PCR process introduces substantial amplification heterogeneity, independent of primer sequence and bulk experimental conditions. This heterogeneity can be attributed both to inherited differences between different template DNA molecules,

and to the inherent stochasticity of the PCR process. The results demonstrate that PCR heterogeneity arises even when reaction and substrate conditions are kept as constant as possible, and therefore single molecule barcoding is essential in order to derive reliable quantitative results from any protocol combining PCR with HTS, such as those frequently used in TCR sequencing studies.

2.1 Introduction

The efficiency of a PCR reaction is known to vary widely, depending on many different factors. These include the properties of the primers [119, 83, 116], the sequence to be amplified [7], in particular its GC content [38, 4], as well as the reaction conditions and type of polymerase. If one wishes to quantify the amount of a given template by PCR ('qPCR') the general approach is to compare an unknown sample to a dilution series of standards, on the assumption that all variables remain the same between sample and standard and hence PCR efficiency remains constant.

The introduction of high throughput sequencing [15, 94], in which many DNA molecules are sequenced individually in parallel, allows the possibility of quantifying many initial target molecules simultaneously by counting the number of times the sequence for each molecule occurs in the output from a sequence run. This approach forms the basis for RNA-seq, in which transcript abundance is measured by sequencing cDNA libraries, and counting the number of sequences mapping to each transcript. Particularly relevant to this thesis, an extension of this approach is the analysis of the antigen-specific receptor repertoire by sequencing cDNA or genomic samples of B or T lymphocytes, and counting the number of times each different receptor is identified [157]. Most current parallel sequencing technologies require nanomolar amounts of starting material (typically $> 10^{10}$ molecules), even when the output of the reaction may only be in the order of 10^7 molecules. In order to achieve this amount of starting material some degree of PCR amplification is usually required. This is especially true when the amount of starting material may be extremely small, for example in the case of single cell RNA-seq [110]. The reproducibility of the PCR amplification process therefore becomes a key factor for accurate quantification.

The use of molecular barcodes provides one approach to dealing with single molecule quantification and mitigating the effects of PCR heterogeneity. A library of diverse short DNA sequences (called 'barcodes', 'unique molecular identifiers' or 'tags' in different studies) are introduced into the molecules to be analysed at an early step in the protocol, in such a way that each target molecule incorporates a different barcode which remains associated with it throughout the amplification protocol. The barcodes can be introduced during a reverse transcription step, or by ligation. For instance, Miner et al [104] and McCloskey et al [98] both ligate nucleotide sequences to uniquely label initial DNA target molecules to identify sequencing redundancy as well as using batch stamps to identify sequencing contamination from other samples. In the work of Casbon et al [32] degenerate base regions are ligated to each DNA fragment to assess whether observed differences between sequence reads are true variants or sequence error. Kivioja et al [77] apply a similar unique molecular identifier technique to human karyotyping. Mamedov et al [93] and Shugay et al [133] use barcodes to provide PCR and sequencing error correction of TCR repertoires.

In this chapter molecular barcoding is used to investigate the extent of variation in PCR amplification on a single molecule basis. In order to rigorously assess the possible sources of this heterogeneity a PCR simulator is developed, which incorporates both amplification and sampling heterogeneity, with which to compare experimental results. The PCR amplification is an example of a branching process, and although there has been considerable theoretical work on such processes, the complexity of heterogeneous branching processes makes analytical modelling challenging in most realistic examples [68, 84, 56]. Detailed models of the physical parameters involved in the PCR cycles have been developed, to answer questions about the probability of replication in an individual cycle or the evolution of the population over a number of cycles [139, 158, 1, 49, 34]. Additionally mathematical models have been used to investigate the error profile in PCR protocols [118] or the presence of non-targeted product through non-specific priming [127]. An increase in computing power has made it feasible to develop PCR simulations using realistic numbers of starting molecules, with reasonable run times. The model we describe includes both an amplification step and a sampling step to simulate the typical workflow of an RNA-seq or repertoire sequencing experiment.

These computational tools can distinguish heterogeneity which derives simply from the sampling process itself (modelled as a zero truncated Poisson process) from stochastic variation in each step of the PCR reaction and inherited variation which may arise from differences between different DNA molecules within the reaction. Our study therefore highlights the potential pitfalls in quantitative analysis of DNA or RNA abundance involving a PCR amplification step, and provides a computational framework which can be used to analyse barcoded PCR data, and identify and quantify the sources of heterogeneity.

2.2 Materials and Methods

All wet lab work was performed by Theres Oakes and James Heather.

2.2.1 Sample collection and processing

Ethics

This study was approved by the joint UCL/University College London Hospitals NHS Trust Human Research Ethics Committee and was carried out in accordance with relevant guidelines and regulations. Written informed consent was obtained from all participants (University College Hospital 06/Q0502/92).

Sample collection

5ml of healthy adult volunteer blood was drawn into Tempus Blood RNA tubes (Life Technologies) and RNA was extracted using the Tempus RNA isolation kit (Life Technologies). Residual DNA was removed using the TURBO DNase kit, and globin mRNA was depleted using GLOBINclear (both Life Technologies).

The KT2 T cell clone was a gift of Prof. A. Lanzavecchia (Institute for Research in Biomedicine, Bellinzona, Switzerland). The clone was grown as described [40]. RNA was isolated using the RNeasy Mini Kit (Qiagen). RNA was treated with RQ1 DNase (Promega) following manufacturers instructions to remove any residual genomic DNA.

Two different protocols were used to amplify and then sequence the T cell receptor chains. All primers are from Sigma-Aldrich and sequences can be found in Table 2.1.

Protocol using single strand ligation (Protocol A)

The DNAse treated RNA was reverse transcribed using oligos complimentary to the 5' region of the TCR constant regions TRAC and TRBC (αRC2 and βRC2, Table 2.1). The mastermix for the reverse transcription was added to the RNA in two stages (molarities for both mastermixes relate to the final volume of 30μl). 11µl of DNase treated RNA were mixed with 0.5µM αRC2, 0.5µM βRC2 and 0.5mM of each dNTP (Invitrogen) to total 19.5µl, and then incubated at 65°c for 5 min and cooled rapidly on ice for $>$ 1 min. 1$\times$ FS buffer (Invitrogen), 5mM DTT (Invitrogen), 30-60 units RNasin Ribonuclease Inhibitor (Promega) and 300 units SuperScript III reverse transcriptase (Life Technologies) were added before incubation at 55°c for 30 min in a total volume of 30µl. 40mM NaOH were added to remove any remaining RNA and the sample was incubated at 70°c for 15 min. 0.5M sodium acetate were added to adjust the pH before the cDNA reverse transcription product was purified using MinElute columns (Qiagen).

The single stranded cDNA was ligated, using T4 RNA ligase (NEB) to a 5' phosphorylated 3' blocked oligonucleotide (T4DNA_6N_SP2, Table 2.1) containing 6 base pairs of random nucleotide barcode and the Illumina sequencing primer SP2. 5µl of cDNA were mixed with 1$\times$ T4 RNA ligase buffer (NEB), 1mM hexammine cobalt chloride, 1.5µM BSA, 0.33mM ATP (NEB), 0.33µM ligation oligo and 20 units T4 RNA ligase 1 (NEB). The ligation was carried out at 16°C for 23 hours followed by a 10 minute heat inactivation step at 65°C. 70µl water were added to the ligation mix before samples were purified at a 1:1 ratio with AMPure XP SPRI beads (BeckmanCoulter) following manufacturers instructions and eluted in 30-35µl water. A second strand was then synthesised, priming from the ligated SP2 sequence. The AMPure bead purified ligation product was incubated with 1$\times$ HF buffer, 0.5µM SP2 primer, 0.5mM of dNTPs and 1 unit of Phusion polymerase in a 50µl reaction at 98°c for 3 min, lowered slowly (1°c/sec) to 80°c, held at 80 °c for 10 sec, lowered slowly (1°c/sec) to 58°c and held at 58°c for 30 sec. After the final extension at 72°c for 1 min, the product was again purified on AMPure beads.

An additional random six base pair barcode was added to the 3' end with a third strand synthesis. The conditions for third strand synthesis were identical to second strand, but using an oligonucleotide complementary to the constant region, and an extension containing the random barcode, and the Illumina SP1 sequencing primer (SP1-6N-I-X-αRC1, or SP1-6N-I-X-βRC1, Table 2.1). A diagram showing the structure of the DNA at this point is shown in Fig 2.1a (top).

The barcoded TCR samples were then amplified in two different consecutive PCR reactions. In the first PCR the P5 and P7 adapters required for Illumina sequencing and an index for multiplex sequencing were added with primers P5-SP1 and P7-LX (Table 2.1). The PCR conditions used were 1$\times$ HF buffer, 0.5µM P5-SP1, 0.5µM P7-LX, 0.5mM dNTPs and 1 unit Phusion; initial cycle: 98°c for 3 min, slowly ramped to 69°c for 15 sec and 1 min at 72°c; cycle 2-4: 98°c for 10 sec and 72°c for 1 min; final cycle: 72°c for 5 min. After bead purification the samples were amplified in a second PCR (1x HF buffer, 0.5µM P5s (Table 2.1), 0.5µM P7 (Table 2.1), 0.5mM dNTPs and 1 unit of Phusion); initial cycle 98°c for 3 min; cycle 1-24: 98°c for 10 sec, 69°c for 15 sec, 72°c for 40 sec; final cycle: 72°c for 5 min. PCR2 products were bead purified and eluted in 30µl water.

Protocol A is represented schematically in Figure A.5.

Protocol using fixed V region primer (Protocol B)

DNAse treated RNA isolated from the KT2 clone was reverse transcribed using oligonucleotides complementary to the 5' region of the TCRβ constant region TRBC. The oligonucleotides also contained a random 12 base pair barcode, the SP1 Illumina sequencing primer and an index for multiplexing. We used two different indices, and each index was placed either next to the SP1 primer sequence (thus providing a spacer between primer and random barcode; SP1-12N-IX-βRC1.1, Table 2.1; Protocol B(i)), or adjacent to the constant region sequence (SP1-IX-12N-βRC1.1, Table 2.1; Protocol B(ii)). Reverse transcription was carried out as in Protocol A.

The cDNA was amplified using a V region specific primer KT2 (VBKT2_1, Table 2.1) and an oligonucleotide complimentary to the Illumina Sequencing Primer SP1. PCR

conditions were 1× HF buffer, 2.5µM primers, 0.5mM dNTPs and 1 unit of Phusion; initial cycle 98°c for 3 min; cycle 1-24: 98°c for 10 sec, 69°c for 15 sec, 72°c for 40 sec; final cycle: 72°c for 5 min. PCR products were bead purified and eluted in 30µl water. The P5, P7 and multiplex index were added in 4 additional rounds of PCR as described for Protocol A.

Library sequencing

Final amplicon products from all sample types were quantified on a Qubit fluorometer (Life Technologies) and sized on a Bioanalyzer (Agilent). Up to 12 samples (at a concentration of 4nM) were multiplexed and sequenced on an Illumina MiSeq, using version 2 chemistry 2x250PE kits.

2.2.2 Data analysis

The FASTQ files produced on the MiSeq were demultiplexed based on the indices added through PCR and analysed using a modified version of Decombinator [145]. Decombinator categorises each TCR sequence read by identifying its constituent V gene and J gene, along with the number of nucleotide deletions from each and nucleotides between V and J regions. The five-part Decombinator classifier ('DCR') is then given by: V gene used, J gene used, number of V deletions, number of J deletions, junctional nucleotide string. The modified version of Decombinator used in this chapter outputs the DCR along with information about the random nucleotide barcode and sequence quality in each sequence read.

For analysis of polyclonal TCR sequence data, the Decombinator output is then passed into a PCR- and sequencing-error correction script. This script first filters sequence reads to remove those where the barcode or sequence quality are poor. It then collects all sequence reads according to their barcode, grouping together those DCRs that appear with identical barcodes. If more than one distinct DCR appears with the same barcode, the DCR with the most copies is taken to be the true sequence of the initial target molecule with that barcode. The others are aggregated into the largest DCR if they are clearly the product of sequencing error, and are discarded otherwise. Next,

Primer name	Primer sequence
αRC1	ACGGCAGGGTCAGGGTTCTGGATAT
βRC1.1	GGTGGGAACACCTTGTTCAGGTCCTC
βRC1.2	GGTGGGAACACGTTTTTCAGGTCCTC
αRC2	GAGTCTCTCAGCTGGTACACG
βRC2	ACACAGCGACCTCGGGTGGGAA
SP1	ACACTCTTTCCCTACACGACGCTCTTCCGATCT
SP2	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
P5	AATGATACGGCGACCACCGAGATC
P7	CAAGCAGAAGACGGCATACGAGAT
VBKT2_1	CTTGGCTATGTGGTCCTTTGC
T4DNA_6N_SP2	[Phos]NNNNNNAGATCGGAAGAGCACACGTCTG-AACTCCAGTCAC[SpcC3]
SP1-6N-I-X-αRC1	ACACTCTTTCCCTACACGACGCTCTTCCGATCT-NNNNNNXXXXXXACGGCAGGGTCAGGGTTCTGGATAT
SP1-6N-I-X-βRC1	ACACTCTTTCCCTACACGACGCTCTTCCGATCT-NNNNNNXXXXXXGGTGGGAACACC(G)TTG(T)TT-CAGGTCCTC
P5-SP1	AATGATACGGCGACCACCGAGATCTACACTCTTT-CCCTACACGACGCTCTTCC
P7-LX	CAAGCAGAAGACGGCATACGAGATXXXXXXGTGA-CTGGAGTTCAGACGTGTGCTCTTCCGATC
SP1-IX-12N-βRC1.1	ACACTCTTTCCCTACACGACGCTCTTCCGATCT-XXXXXXNNNNTNNNNTNNNGGTGGGAACACCTT-GTTCAGGTCCTC
SP1-12N-IX-βRC1.1	ACACTCTTTCCCTACACGACGCTCTTCCGATCT-NNNNTNNNNTNNNNXXXXXXGGTGGGAACACCT-TGTTCAGGTCCTC

Table 2.1: Sequences of primers used in protocols A and B, where X represents a variable but known nucleotide (e.g. for sample indexing) and N represents a random unknown nucleotide (for molecule barcoding)

the set of different barcodes associated with the same DCR is considered. Barcodes that are similar and are observed in the context of the same DCR are considered to be derived from the same initial molecule and are therefore aggregated. The size of the set of distinct barcodes found in the context of the same DCR provides us with a measure of the number of initial copies of that T cell receptor present in our sample (the clone size). For this study, we additionally count the number of copies of each barcode-DCR combination (the barcode family size) to provide us with information about the amplification of the initial molecules.

The structure of the available barcode pool is inferred from the distribution of the

number of times each barcode is found to have labelled a different cDNA molecule (barcode-labelling events) across all experiments in this study. The barcode-labelling events data are fitted by various zero-truncated mixed Poisson models using custom functions (Appendix A), minimised using the Optimise function of SciPy in Python. The parameters of the fitted models are used to infer the structure of the pool of available barcodes.

From the healthy volunteer PBMC samples, a median of 688,001 (minimum 386,904, maximum 1,041,485) sequence reads contain an identifiable TCR, and after clustering by bacodes this is reduced to a median of 20,700 (minimum 9,770, maximum 50,253) total initial molecules that are observed in the sequencer output. There are a median of 13,629 (minimum 7,134, maximum 30,614) distinct TCRβs in these samples.

2.2.3 PCR simulator

Simulation of labelling, amplification and sequencing of samples of molecules is performed with functions written in Python which are available in Appendix B. Briefly, at each cycle a molecule has a chance to successfully replicate. The probability of successful replication is determined by the PCR model chosen. If replication is successful, nucleotide error is incorporated at a given rate by choosing at random whether a given position in the sequence contains error and if so which nucleotide is incorporated incorrectly. Molecules to be sequenced are selected at random from the amplified pool and sequencing error is incorporated into these molecules similarly.

In this simulation, PCR is treated as a type of branching process. A general branching process models a population where each individual produces a number $(0,1,2,\ldots)$ of offspring after each generation. Each element of the population behaves identically, that is, the probability of producing a given number of offspring is distributed identically for all elements in each generation. For PCR, where a molecule of DNA can be replicated, degraded or neither in each thermo-cycle we have the number of possible offspring from each molecule of DNA being 0, 1 or 2 after each discrete generation of a fixed time (i.e. the cycle time) for all molecules. A discussion of branching processes in the context of PCR can be found in [75]. In the work presented in this chapter, a number

of variations of the PCR simulator are used. In the most basic model implementation, the number of descendents from a molecule in one cycle is equal to 2 with probability p and is equal to 1 with probability $1-p$. p is referred to as the efficiency of the PCR reaction, which remains constant in the basic model across cycles and between molecules. In other implementations of the model, described in the results section, molecule degradation is possible (allowing for 0 offspring) or the efficiency is non-constant in time. Additionally, we implement a model where the efficiency for each molecule is selected from a distribution, either before the start of the PCR reaction and inherited from ancestor to descendant molecule (creating a separate branching process for each of the initial target molecules), or at each cycle for each molecule (disrupting the property of branching processes that each item in the population behaves in an identically distributed fashion).

2.3 Results

2.3.1 Heterogeneous amplification efficiency demonstrated by unique molecular barcoding of cDNA molecules.

We reverse transcribed a sample of TCR RNA from peripheral blood T cells and then ligated a primer that contained a unique barcode followed by a sequence corresponding to the Illumina SP2 sequencing primer (Protocol A). The individually tagged mixtures of different TCRα and TCRβ chains were amplified using constant region 3' primers and a 5' primer homologous to the Illumina SP2 sequence on the ligated oligonucleotide (Figure 2.1a, top). The resulting amplified PCR reaction was diluted and sequenced using the standard Illumina protocol (illustrated diagrammatically in Figure 2.1b). The number of times each barcode was present in the sequence data was then counted. We refer to all sequences that have an identical barcode as a barcode family, and refer to the number of molecules present with this barcode as a barcode family size. Although each cDNA molecule was ligated to a different barcode, and the starting frequency of each barcode should then be uniform and independent of the clone size of the TCR sequence with which it was associated, the observed distribution of barcode family sizes in a polyclonal sample was very heterogeneous (Figure 2.2a,

top). Thus, while the majority of barcode families were of size one, some barcodes occurred over 100 times. A similar pattern was observed for TCRα and TCRβ sequences, indicating that the heterogeneity was not some special feature of the sequence being amplified. We repeated this analysis on different polyclonal samples, sequenced at different depths (Figure 2.2a, middle) and with different numbers of observed barcodes (and therefore different numbers of initial target molecules) carried through the protocol (Figure 2.2a, bottom). Extensive heterogeneity, varying over two orders of magnitude, was observed in each case. Without barcoding, this heterogeneity would have a substantial impact on analysis of both the diversity and the structure of the TCR repertoire (Figure A.2).

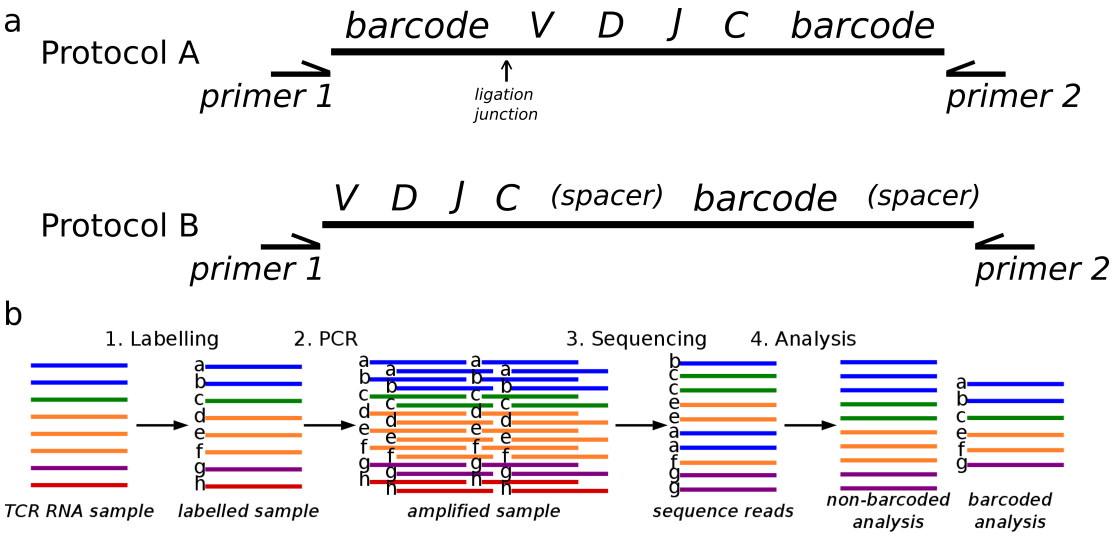

Figure 2.1: Schematic of the PCR amplification study

(a) Schematic of the target TCR molecule in Protocols A and B showing the position of the barcode for molecular identification and the PCR priming sites. In the T cell receptor portion of the molecule, V, D, J and C refer to the Variable, Diversity (β chain only), Joining and Constant regions of the TCR α or β chains (not to scale). The two alternative possible positions for the barcodes in protocol B(i) and B(ii) are shown in brackets. The Illumina sequencing primers, indices to allow for multiplexing of samples, and the Illumina adaptor sequences are not shown.
(b) Schematic of experimental and computational protocol used to sequence and analyse TCRs from isolated RNA. Barcodes (represented here by lower case letters) are included in each TCR molecule together with a known sequence (SP2). PCR is then performed to amplify the sample. The amplified pool of molecules is diluted and introduced to the sequencer, where a sample of molecules will adhere to the flow cell and be sequenced. Repertoire analysis is performed on the sequencing data, with the barcodes allowing correction of biased PCR amplification as well as correction of sequencing errors.

One possible explanation for the observed distribution was the heterogeneous template

mixture of cDNAs due to the diversity of the TCR repertoire. Although the primers and the primer binding regions were the same for all amplified molecules, the intervening sequences were heterogeneous since they represented many different TCR sequences. Thus, heterogeneous amplification could reflect differences in target replication by polymerase. In order to simplify the experimental model, and limit the heterogeneity arising from using a complex pool of substrate molecules (a natural TCR repertoire), we labelled and amplified a TCR sequence (α and β chain) from a human T cell clone, KT2, which expresses only one T cell receptor. As predicted, the vast majority of sequences from these samples were identical (Figure A.1). To our surprise the distribution of barcode frequencies was still just as heterogeneous (Figure 2.2b, top). Thus even under conditions where we were amplifying a single target (namely the KT2 TCR α or β chain), and primer and reaction conditions were identical for all amplified molecules, we still observed a difference of two orders of magnitude in the number of molecules derived from single starting template cDNA molecules.

We considered two further possible sources of heterogeneity which could potentially contribute to the observed range of barcode family sizes. The first was the single-stranded DNA ligation step used in Protocol A (Figure 2.1a, top). Although this allows a single primer to be used for a heterogeneous mixture of DNAs and avoids the need for complex primer multiplexing, it creates a potential for heterogeneity at the end of the cDNA template molecule as a result of incomplete reverse transcription of the RNA. A second possible cause of heterogeneity are the barcodes themselves. In particular Pan et al [116] have shown that the basepairs immediately adjacent to the PCR primer can have a small effect on amplification efficiency.

In order to address the first issue, we performed a further PCR using a fixed primer within the V region of the KT2 TCRβ chain instead of the single stranded ligation step (Protocol B, Figure 2.1a, bottom). The unique barcodes were introduced during the reverse transcription step, and were placed either adjacent to the primer as previously (Protocol B(i)), or separated from the primer by a six base pair index region (Protocol B(ii)). The results of these further sets of PCR are shown in Figure 2.2b (middle and bottom panels). The omission of the ligation step (Protocol B) decreased the amount of heterogeneity, although differences in amplification of greater than 10 fold remained.

However, Protocol B is not ideal for TCR repertoire sequencing studies because the use of V-gene specific primers introduces additional bias as well as possibly preventing non-canonical rearrangements from being observed.

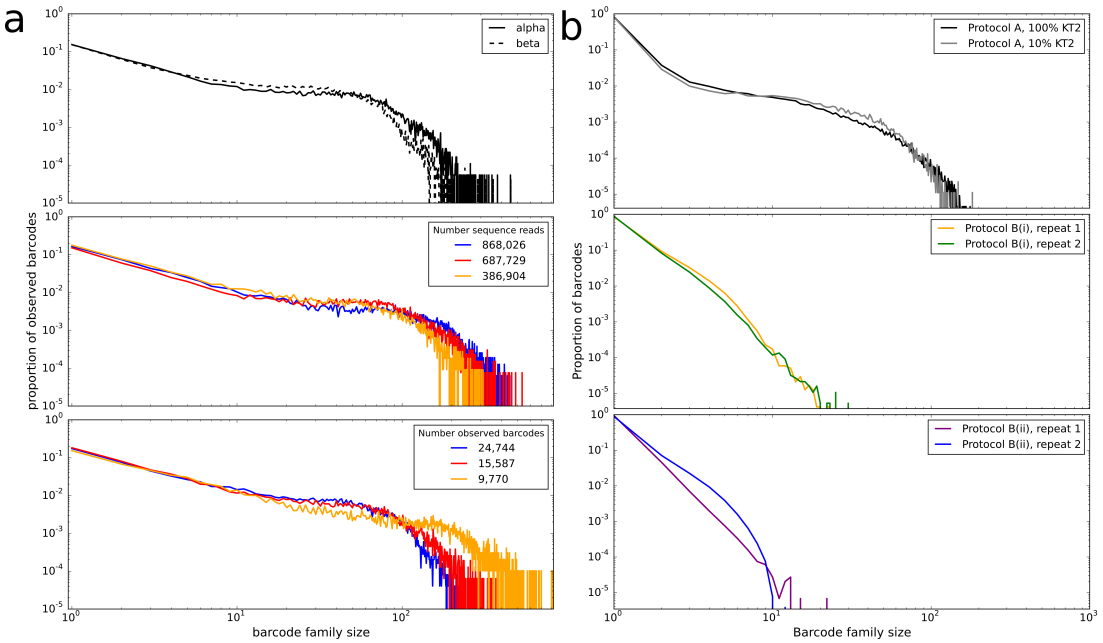

Figure 2.2: Long-tailed distribution of barcode family sizes observed.
(a) The distribution of observed barcode family size (the number of sequence reads occurring in the sequencer output that originate from the same initial target molecule in the sample) in polyclonal TCR sequence data (Protocol A) from healthy volunteer T cells. Upper: TCR α chain (solid line) and β chain (dotted line) data. Middle: TCR repertoires sequenced at different depths. Bottom: TCR repertoires with different numbers of observed barcodes, representing the number of initial molecules.
(b) The observed barcode family size distribution observed in TCR sequence data from a sample of RNA isolated from a T cell clone (KT2, responding to tetanus toxoid [40]). Upper: TCR α chain (solid line) and β chain (dashed line) from protocol A. Middle: TCR β chain data from protocol B, using the oligonucleotide with 6bp spacer between the sequencing primer and the barcode. Bottom: TCR β chain data from protocol B, using the oligonucleotide with the barcode directly next to the sequencing primer.)

2.3.2 Barcode family size is not dependent on barcode sequence, barcode clash or non-uniform barcode primer frequencies

The heterogeneous amplification observed could hypothetically be caused by the barcode itself since the polymerase must amplify the barcode in each cycle. To investigate this, we first considered whether barcodes that appear more amplified have a tendency

to contain more or fewer G or C nucleotides (Figure 2.3a). However, there was no obvious relationship between the frequency of particular barcodes and their GC content. Furthermore, the frequency rank of the same barcode in any two different sequence runs was uncorrelated (Figure 2.3b). A high barcode family size did not therefore appear to be the result of a particular barcode sequence or sequence motif. Additionally, to account for the fact that the amplification effect might be to do with relative, rather than absolute, barcode 'fitness', we considered all pairs of barcodes that are both observed in any pair of experiments. If the amplification was determined by the barcode we might expect, for example, that if barcode A is larger than barcode B in experiment 1 then it would also be larger in experiment 2. We found no correlation between the frequencies of any two barcodes that appear together in a pair of experiments (Figure 2.3c), implying that the barcode sequence itself does not determine the efficiency with which each molecule is amplified. We also examined whether the observed barcode family size might be an artefact introduced during the sequencing reactions, perhaps by heterogeneity in bridge PCR on the flow cell. If this were the case we might expect that molecules from large barcode families are located in close proximity on the flow cell. However there was no observable relationship between barcode family size and location of molecules on the flow cell (one representative frame shown in Figure 2.3d).

The barcodes used in these experiments should theoretically contain randomly chosen nucleotides at each of the 12 positions, giving a total of $4^{12} \approx 1.7 \times 10^7$ possible barcodes, each appearing an equal number of times. In practice, the methods of oligonucleotide synthesis likely result in slightly different incorporation efficiencies of different nucleotides at each position [54]. In addition, the number of target molecules barcoded in our T cell samples is often within an order of magnitude of the number of available barcodes, resulting in a significant probability that the same barcode is used more than once ('barcode clash') (Figure 2.4a). In order to assess the impact that this barcode clash might have on the observed barcode family sizes, we first simulated barcoding molecules from a large, uniformly distributed pool of available barcodes and measured the proportion of molecules that were uniquely barcoded (Figure 2.4b). This value depends on the ratio of the number of available barcodes (size of the barcode pool) to the number of molecules to be barcoded. In these simulations we also mea-

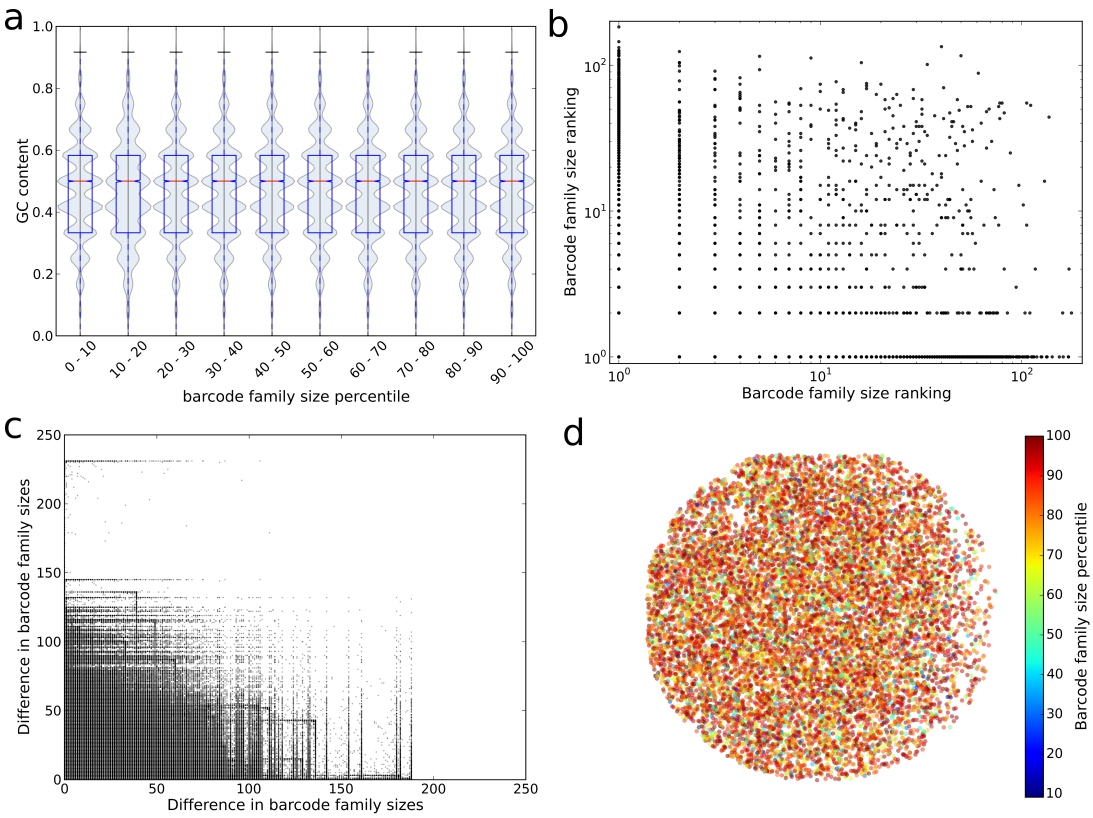

Figure 2.3: Final barcode family size is unrelated to properties of the sequence being amplified

(a) Distribution of GC content of the 12-nucleotide random barcodes by barcode family size percentile. Data from a sequence run of healthy volunteer PBMC TCRs.
(b) The correlation between the barcode family size ranking in any pair of runs for those barcodes that occur in more than one of the eight monoclonal KT2 TCR sequencing runs (Protocol A) in this study (R-squared < 0.0003). Ranking is ascending, and barcodes that have the same family size in a run are given the same ranking. There is no gap introduced in rankings when more than one barcode occupies a particular ranking, as such for small barcode family sizes ranking is equivalent to barcode family size.
(c) For those pairs of barcodes that appear together in any pair of the eight KT2 sequencing runs (Protocol A) in this study, the relationship between the difference in barcode family sizes in one run and in the other. R-squared < 0.0004.
(d) Position of TCR molecules on the flowcell, coloured by barcode family size percentile. Representative example of a single frame from one flow cell from a sequencing run in this study.

sure the maximum observed barcode clash size (Figure 2.4c), which in contrast also depends on the absolute number of available barcodes and molecules to be barcoded. These simulations show that in our protocol (barcoding in the order of 10^6 molecules with 10^7 available barcodes) around 90% of molecules get a unique barcode and the maximum clash size is predicted to be below 4. Thus barcode clash is unable to account for the range in barcode family sizes we observe in our data.

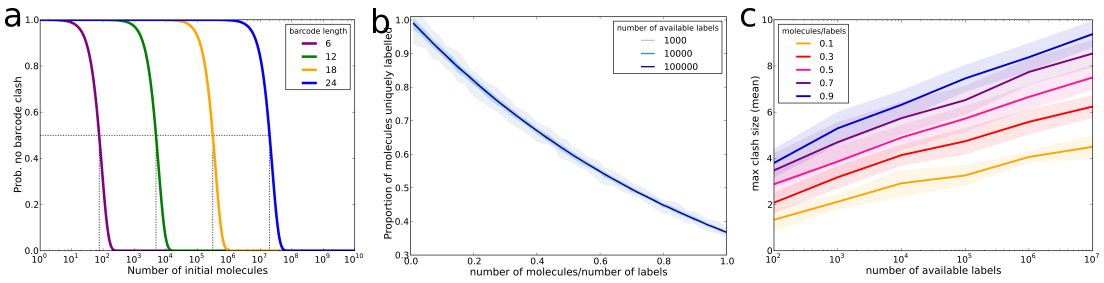

Figure 2.4: Barcode clashes do not explain the observed PCR amplification heterogeneity
(a) The probability that no two molecules receive the same label (barcode clash) when initial molecules are labelled with a pool of random nucleotide barcodes of the indicated length. The dotted lines indicate the number of molecules that can be labelled with a 50% chance of no barcode clash occurring.
(b) The proportion of initial molecules that receive a unique barcode when barcoding is simulated with the indicated number of available barcodes, uniformly distributed. The number of molecules to be barcoded is expressed as a proportion of the number of available barcodes. Data shown is the mean and standard deviation of 50 repeated simulations.
(c) The maximum number of initial molecules that receive the same barcode when barcoding is simulated with the indicated number of available barcodes that are uniformly distributed. The number of molecules being barcoded is indicated by colour, expressed as a proportion of the number of available barcodes. Data shown is the mean and standard deviation of 50 repeated simulations.

It is likely that the pool of barcodes we have available for labelling is not exactly uniformly distributed, which could lead to increased barcode clash. We simulated the barcoding, amplification and sequencing protocol using normally or lognormally distributed barcode frequency distributions, but this had little effect on the observed barcode family size distributions when compared to uniquely barcoding every molecule or to the expected distribution if every initial molecule was represented equally in the post-PCR amplified pool (Figure 2.5a). We also derived the empirical distribution of barcodes in our initial oligonucleotide pool (Appendix A) and Figure A.3) and simulations using this distribution do not show a barcode family size distribution deviating far from the sampling distribution expected from a uniformly distributed amplified pool

(Figure 2.5b). The output of the barcoding, amplification and sequencing pipeline is therefore robust to the likely occurrence of barcode clash and non-uniform barcode frequencies.

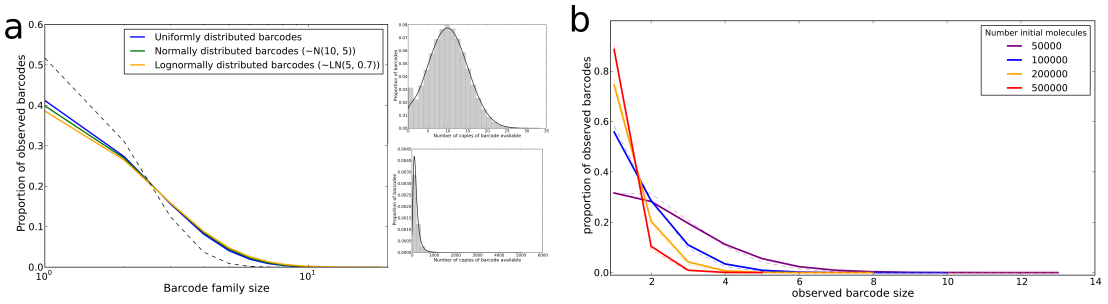

Figure 2.5: Non-uniform barcode availability does not explain observed PCR amplification heterogeneity

(a) The observed family size distribution after simulation of barcoding 250,000 initial molecules from uniformly or non-uniformly distributed pools of 500,000 available barcodes, 10 cycles of PCR at efficiency 0.5 and sequencing 300,000 molecules from the amplified pool. The distribution of available barcodes for the non-uniform simulations is shown in the inset (green: normal distribution (restricted to values > 0), orange: lognormal distribution). Data shown are the mean and standard deviation of 10 repeated simulations. The grey dotted line shows the barcode family size distribution that would be expected if the molecules to be sequenced were drawn from a uniformly distributed amplified pool, in which all molecules had been uniquely barcoded and amplified equally.

(b) The observed barcode family size distribution when the indicated numbers of initial molecules are barcoded from a pool of 4^{12} potential barcodes with barcode availability distributed as predicted from empirical labelling events observed (details in Appendix A with the empirical distribution shown in Figure A.3). PCR cycles (25 cycles, 0.75 efficiency) are simulated on the labelled molecules, and samples of size 100,000 are selected from the amplified pool. The solid line represents the mean of 10 repeated simulations. The dashed line shows the expected distribution had the sample been drawn from a uniformly distributed amplified pool, in which every initial molecule had been barcoded uniquely and amplified by the same amount.

2.3.3 Inherited differences in PCR efficiency are necessary to explain the observed diversity in barcode family size.

The experimental pipeline involves amplification followed by subsampling for sequencing, which can introduce Poisson non-uniformity even when the amplified pool of barcoded molecules is uniform. Furthermore, PCR efficiencies of less than 100% can introduce non-uniformity resulting from the inherent stochasticity of the PCR process [117]. In order to examine how variable efficiency and sampling could affect

observed barcode family size distributions a PCR simulator was developed in which molecules are barcoded, amplified and then sampled in silico. The simulator is outlined schematically in Figure 2.6a. In its most basic implementation (modelling PCR as a straightforward branching process with no error) the simulator can perform a full simulation (labelling initial molecules, performing 15 PCR cycles with efficiency 0.8, sampling and sequencing including sequencing error) on 10^5 initial molecules in approximately 12 seconds on a standard specification laptop (Figure 2.6b). Introducing PCR error substantially increases the simulation time, although altering the error rate further does not alter simulation time. Use of parallelisation and cluster research computing platforms would make PCR simulation including error of large numbers of initial molecules feasible.

The simulated barcode distributions (the number of molecules present after amplification that are derived from each initial molecule) at different efficiencies are shown in Figure 2.6c. The introduction of less than 100% efficiency introduces some barcode family size heterogeneity (in the amplified pool, before sampling for sequencing) as described previously [117]. This variation arises because, in every replication cycle, any individual molecule may or may not replicate with a probability determined by the overall efficiency. The substantial shoulder observed in the distributions correspond to molecules which fail to be replicated in the first cycle of PCR and hence are present at half the average number of copies. However, the heterogeneity caused by low efficiencies is averaged out over many molecules and the majority of barcode family sizes are within a factor of two of each other at the end of the PCR reaction.

When a sample of molecules is drawn at random from the amplified pool (to simulate the process by which molecules from the amplified sample are diluted and introduced to the flow cell to anneal to complementary capture oligonucleotides), the observed barcode family size is further diversified depending on the ratio of number of sequenced molecules to number of initial molecules (the ‘sample ratio’, Figure 2.7a). These observed barcode family size distributions follow a Poisson distribution (as an approximation to a binomial distribution), scaled to account for the fact that we cannot count those barcodes with an observed family size of zero (a zero-truncated Poisson). The Poisson distribution is the expected distribution when sampling from an amplified pool in which

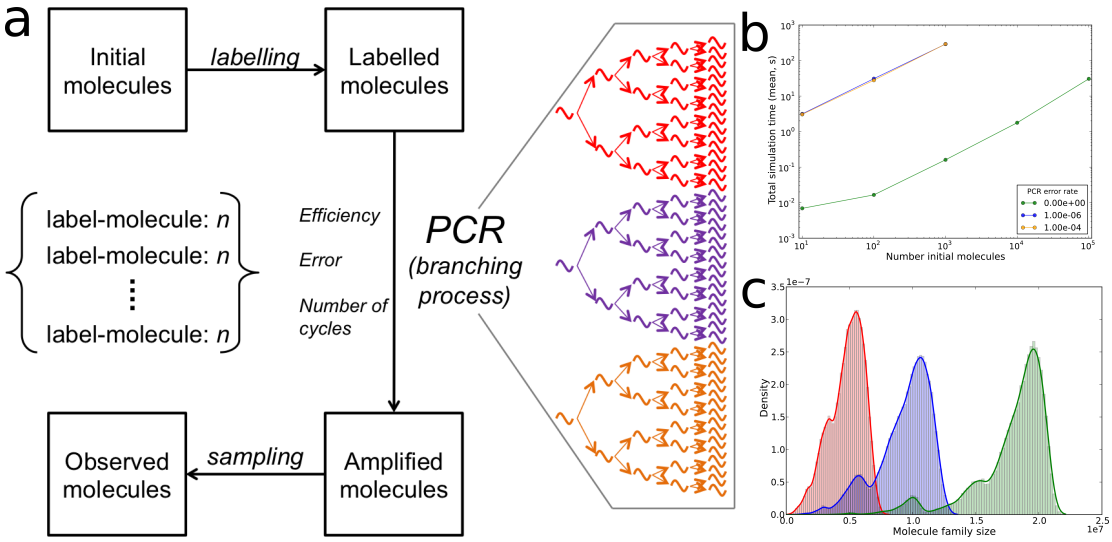

Figure 2.6: PCR simulator software

(a) Schematic of the PCR simulator software used in this study. The software includes adding barcodes to molecules (labelling), PCR amplification with a specified number of cycles, efficiency model and error rate, and sampling and sequencing from the amplified pool.

(b) Time taken to perform a full simulation, which includes initialisation, labelling initial molecules, PCR cycles (using a standard branching process model), sampling from the amplified pool and sequencing. Simulations are performed with the indicated PCR error rate (per base per cycle) and the given number of initial template molecules. Simulations consist of 15 cycles of PCR with efficiency 0.8, a sample size equal to the number of initial molecules being chosen from the amplified pool and sequencing with error rate 10^{-4}. Data shown is the mean of 5 repeated simulations at each set of conditions, as measured on a 2.8 GHz Intel Core i7 MacBook Pro.

(c) The distribution of the number of copies of each of 100,000 initial target molecules after 25 cycles of PCR at efficiencies of 0.85 (red), 0.9 (blue) or 0.95 (green).

each barcode is present the same number of times. If the PCR process in our experiments behaved as a straightforward branching process we would expect our experimental observed barcode family size distributions to also follow a zero-truncated Poisson distribution, with the Poisson parameter providing information about how many initial molecules there were in our sample. However, it can be seen that our data from Protocol A does not belong to the same distribution family as the simulated distributions (Figure 2.7a), suggesting that these samples were not drawn from a uniformly distributed post-PCR pool and that neither heterogeneity resulting from a low PCR efficiency or from the sampling process can account for the broad distribution of barcode family sizes observed.

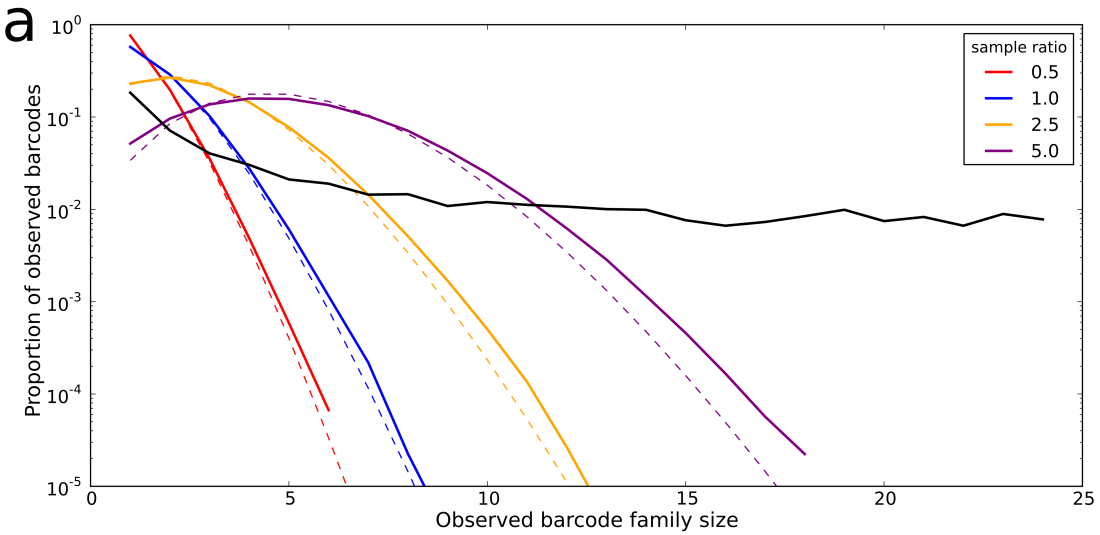

Figure 2.7: Barcode family size distributions with fixed amplification efficiency
(a) The distribution of observed barcode family sizes (coloured lines) after simulating PCR cycles (25 cycles at 0.9 efficiency) on 100,000 initial molecules and then sampling from the amplified pool to select those molecules that are observed in the sequencer output. The number of molecules sequenced is expressed as a proportion (the 'sample ratio') of the number of initial molecules (100,000). The solid coloured lines are the mean of 5 repeated simulations, and the dashed coloured lines are the expected distribution (a zero truncated Poisson with parameter equal to the sample ratio) if the sample was drawn from a uniformly distributed pool (which would occur if every initial molecule was uniquely barcoded and amplified identically). The black solid line is a representative example of the barcode family size distribution observed in TCR sequencing data from healthy volunteer PBMC.

We therefore tried to formulate variations of the branching process model of PCR that could explain the broad barcode family size distribution observed. The starting point is a standard branching process model of PCR ('Model 1') where the efficiency of the PCR (between 0 and 1) refers to the probability that a molecule will replicate successfully in a cycle. Using this model, we simulate PCR and sampling, and show that the resulting barcode family size distributions do not diverge significantly from the expected Poisson distribution regardless of the efficiency used (Figure 2.8a). Next, a target degradation model ('Model 2') was used. Model 2 is set up as for Model 1, except that when a molecule fails to duplicate in a cycle there is a chance that it instead degrades and is no longer available to be amplified in later cycles of the PCR. Again, simulation of this model does not reproduce the large deviation from the a Poisson distribution that is seen in our data (Figure A.4a).

Next, we introduce competition for resource, which affects the success rate of dupli-

cation of molecules. This abstract 'resource' covers, for example, the availability of dNTPs and primer in the PCR mixture, and the ability of the enzyme to process the molecules inside the time frame given in the PCR protocol. The first resource competition model ('Model 3') is one in which there is a fixed, constant amount of resource available and the probability that a molecule successfully replicates in a cycle is given by the available resource divided by the number of molecules present at the start of the cycle. As such, the efficiency of the reaction decreases through the cycles once the number of molecules present exceeds the capacity of the available resource to process all those molecules in one cycle. Figure 2.8b shows that this model cannot reproduce the spread of barcode family sizes we observe in the data. An alternative resource competition model ('Model 4') involves degradation of resource as it is used, at a given degradation rate. This model is also unable to account for our observed barcode family size distributions (Figure A.4b).

Instead of a constant efficiency across all molecules and all cycles, we imagine that in a given cycle some molecules are able to replicate more efficiently than others. For instance this variation may depend on the position of the molecule within the sample (which may affect e.g. proximity to primer) or the conformation of the molecule (which may affect ability of the primer to bind). We introduce a variable efficiency model ('Model 5'), where the probability that a given molecule will replicate in a given cycle is chosen from a defined distribution. Model 5 is implemented by selecting efficiencies from a normal distribution with a variety of parameters (Figure 2.8c). A low mean efficiency and a large standard deviation produces the most divergence from the expected barcode family size distribution, and is able to account for the majority of the spread seen in barcode family size observed in the KT2 data from Protocol B. However, none of the parameters investigated was able to reproduce the observed spread of family sizes observed in polyclonal or monoclonal data from Protocol A.

We adapted Model 5 to include the constraint that once an efficiency is chosen for a molecule in cycle 1 this same efficiency is inherited by all molecules produced from this initial molecule ('Model 6'). Simulation of PCR and sampling using Model 6 was performed, and showed that inherited efficiencies could produce a substantial amount of spread in the barcode family size distribution when the efficiency distribution has a

low mean and a relatively large standard deviation (Figure 2.8d). The observed barcode family size distribution from Model 6 can be seen to be broadly comparable to that seen in our experimental data (from Protocol A) for these parameters. In contrast, the distribution which arises from Model 5 is sufficient to account for most of the heterogeneity observed when using a fixed primer instead of ligating a primer to the end of the cDNA (Protocol B).

2.4 Discussion

PCR is a fundamental and ubiquitous tool of molecular biology laboratories. The combination of PCR and HTS, in particular, has driven an explosion in DNA sequence acquisition. In many of these applications, for example RNA-seq and lymphocyte antigen receptor repertoire studies, the quantification of transcripts is critical, since the analysis is often based on counts of specific sequences. The avoidance of PCR bias is therefore critical and much effort has been expended on trying to control and mitigate bias. In this work the consistency of PCR amplification is examined, using molecular barcodes to follow amplification of single molecules. The distribution of the number of copies of an initial molecule observed in sequencer output is found to vary over a wide range, even when primers, target sequence and bulk PCR conditions are kept constant, and in a manner which appears to be independent of barcode sequences.

The differential binding properties of different primers, and secondary structure within target sequences are well-established causes of PCR biases. Multiplex PCRs, for example, frequently show different efficiencies for different primer/target combinations. This bias is a known confounder of T cell repertoire studies, for example. As a result, our lab and others [133] have developed techniques that use various types of 5' RACE, and thus can amplify with amplicon-independent primers. However, the variation in target sequence to be amplified is obviously a variable that cannot be avoided. In this study we therefore consider the extent to which amplification bias can be attributed to sequence variability. We compare the amplification of heterogeneous mixtures of α or β T cell receptor chains (typically containing $> 10^4$ different sequences) with amplification of a monoclonal T cell receptor from a T cell clone (this clone in fact

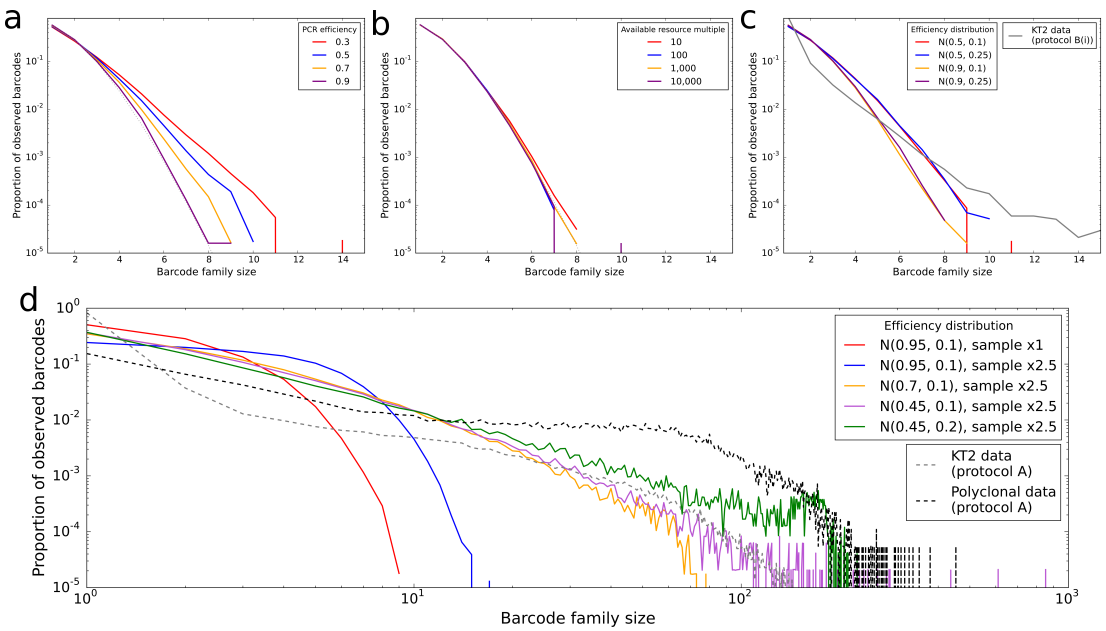

Figure 2.8: Barcode family size distributions under different models of PCR heterogeneity

Observed barcode family size distributions observed under different models of PCR duplication. Simulations performed with 10,000 initial molecules, 25 cycles of PCR (with no error) and sequencing of 10,000 molecules selected from the amplified pool. Simulations were repeated 10 times and the mean and standard deviation are shown. The dotted lines represent the expected distribution if every initial molecule is barcoded uniquely and represented equally in the amplified pool. Grey and black lines represent the distributions observed experimentally in the indicated experiments. Models described in the text but not displayed here can be found in Figure A.4.

(a) Model 1: Standard branching process of PCR, with the indicated efficiencies. The efficiency is the probability that a given molecule will duplicate in a given cycle.

(b) Model 3: Model of PCR where the duplication efficiency depends on competition between target molecules for a constant level of resource, given as a multiple of the number of initial molecules.

(c) Model 5: Variable efficiency model of PCR, where the probability of a given molecule replicating in a given cycles is selected from a normal distribution (restricted to [0, 1]) with the indicated parameters (mean and standard deviation).

(d) Model 6: Inherited efficiency model of PCR, where the probability of replication in a given cycle is identical for all molecules derived from the same initial molecule. The efficiencies for the initial molecules are selected from a normal distribution with the indicated parameters (mean and standard deviation). 25 PCR cycles are simulated on 10,000 initial molecules, and then a sample is drawn from the amplified pool at a multiple of 2.5 times the number of initial molecules. The observed barcode family size distribution shown is the mean of 10 repeated simulations.

expresses more than one TCR chain, a common feature of T cells [36]). Unexpectedly, PCR amplification efficiency (measured by the number of observed molecules derived from a single ancestor) varies broadly, both for the polyclonal and monoclonal populations. Indeed the extent of variability is very similar, suggesting that the actual sequence of the TCR variable region is not the major cause of different amplification rates. Our results do not, of course, imply that all sequences will be amplified equally. Indeed the length of the target and the GC content are well known to influence PCR efficiency [4]. Rather our results suggest that even when amplifying relatively small amplicons ($<$1KB) whose sequences are all rather comparable, substantial variation remains. Some degree of heterogenous amplification of T and B cell receptors has been observed previously [132, 154, 133], although these studies have not focused on analysis of the distribution of the variation, or its relationship to inherent stochasticity of the PCR process.

These data suggest that the sequence of the ligated barcodes is not the cause of the observed differential amplification, since barcode family size is not correlated between experiments. Although previous studies have shown a small effect of sequence variability adjacent to the PCR primer on efficiency [116], we have directly compared placing the random barcode sequences immediately next to the primer, or at a distance of six base pairs, and did not observe any significant difference in heterogeneity. Indeed it seemed a priori unlikely that if the variation cannot be attributed to differences between V region sequences it could be caused by 12 base pair barcodes. Additionally, analysis of the structure of the pool of the random barcodes that are used to label the initial molecules suggests that while there is potential for barcode 'clashes' (where the same barcode sequence is used to label more than one initial molecule), these are not large enough or prevalent enough to be the reason for the large barcode family sizes observed. We do, however, present some theoretical and simulation results that can help to guide the size of barcode pool size in different scenarios. These results suggest that 12 base pair barcodes (providing in the order of 10^7 different barcode sequences) are sufficient to label samples of DNA targets in the order of 10^6 molecules.

The bulk conditions in all the PCR reactions obviously cannot account for the intra-experimental variation. However, as discussed previously, PCR is by its nature a

stochastic process since at each cycle a molecule will be either replicated or not replicated with some probability p, which will be less than 1 for all reactions in which replication efficiency is not 100%. For example, the PCR efficiencies in our model system (which we have measured using qPCR on plasmid dilutions) are typically in the order of 80% - 90%. Furthermore, it is possible that there is local heterogeneity in the PCR vessel itself: for example temperature gradients, or heterogeneity introduced by phase shifts at the plastic/liquid or liquid/gas surfaces. We therefore examined the implications of different models in detail using a branching process PCR simulator.

Simulation demonstrated clearly that lower efficiencies, a range of efficiencies, competition and resource limitation can all introduce some variation in the predicted output of the PCR for different molecules. As might be predicted, the extent of variation increases with cycle number, and with low and more variable efficiencies. The goal of minimising the number of cycles, and maximising efficiency does therefore lower overall expected variance of product molecular counts. However, the extent of the variance due to these properties is limited and does not explain our observed results. The only model we considered that was able to produce substantial variance in output comparable to that observed in Protocol A (Figures 2.2a and b, top) is an inherited efficiency model, where all molecules produced from an initial molecule retain the same efficiency though all cycles. This result, too, is related to well known evolutionary theory where significant divergence can only occur when selection operates on the inherited properties of the individual. A clue to the cause of the observed heterogeneity is provided by the observation that it is reduced, by omitting the single stranded ligation step, and instead using a fixed primer in the V region (Figure 2.2b, middle and bottom panels). Since the length of the cDNA molecules produced may be variable, due to the nature of the target molecules, variation in length or composition of the cDNA at the 3' end may be sufficient to significantly alter PCR amplification efficiency. Thus for TCR or BCR repertoire sequence, both multiplex PCR and RACE protocols have the potential to introduce substantial heterogeneity in amplification efficiency, which will materially affect quantitative features of the observed repertoire, substantially increasing the range of clone sizes observed (Figure A.2). Barcoding therefore becomes essential for accurate quantification of transcript number.

In conclusion, we consider the implications of our findings for the community routinely using PCR for quantitative analysis of RNA or DNA populations. The major lesson is that molecular barcoding provides an essential tool that can mitigate for the effects of PCR heterogeneity. This is especially important for studies whose primary output is the comparative quantification of many diverse nucleotide fragments within a mixture, such as repertoire analysis. In situations where single molecule barcoding is difficult, or not practical, every effort needs to be taken to maximise the efficiency of the PCR reactions and minimise the number of cycles. In the longer term, single molecule amplification-free DNA sequencers, which are currently in development, may remove the requirement for a PCR amplification step altogether. In the meantime, it continues to be important to appreciate the inherent stochasticity of the PCR process, and its possible effects on quantitative aspects of molecular biology.

Chapter 3

Immune tolerance maintained by cooperative interactions between T cells and antigen presenting cells shapes a diverse TCR repertoire

The study in this chapter was undertaken in conjuction with Dr Chris Watkins and is published in [17].

The T cell population in an individual needs to avoid harmful activation by self peptides while maintaining the ability to respond to an unknown set of foreign peptides. This property is acquired by a combination of thymic and extra-thymic mechanisms. We extend current models for the development of self/non-self discrimination to consider the acquisition of self-tolerance as an emergent system level property of the overall T cell receptor repertoire. We propose one way in which tolerance can be established is at the level of the antigen presenting cell/T cell cluster, which facilitates and integrates cooperative interactions between T cells of different specificities. The threshold for self-reactivity is therefore imposed at a population level, and not at the level of the individual T cell/antigen encounter. Mathematically, the model can be formulated as a linear programming optimisation problem that can be implemented as a multiplicative update algorithm, which shows a rapid convergence to a stable state. The model

constrains self-reactivity within a predefined threshold, but maintains the diversity and cross-reactivity which are key characteristics of human T cell immunity. We show further that the size of individual clones in the model repertoire remains heterogeneous, and that new clones can establish themselves even when the repertoire has stabilised. Our study combines the salient features of the 'danger' model of self/non-self discrimination with the concepts of quorum sensing, and extends repertoire generation models to encompass the establishment of tolerance. Furthermore, the dynamic and continuous repertoire reshaping which underlies tolerance in this model suggests opportunities for therapeutic intervention to achieve long-term tolerance following transplantation.

3.1 Introduction

Vertebrate immune system recognition uses antigen receptors produced by stochastic and hence unpredictable molecular recombination events. In this study we propose a new explanation for how the T cell compartment of the immune system may use a stochastic set of receptors, whose specificities are not predetermined, to develop a useful repertoire. The requirements we impose are that the repertoire of antigen receptors should cover the set of non-self antigens as comprehensively as possible, in order to provide robust protection against any potential exposure to infectious pathogens. At the same time the system must remain tolerant to the set of self antigens and generally avoid autoimmunity. The fundamental aspect of our hypothesis is that self/non-self discrimination is an emergent property of the combined population of T cells, and cannot be linked by a one-to-one mapping to the individual binding strength spectrum of individual T cells and their receptors. The model we propose has important implications in the context of transplantation, since it suggests that the repertoire can be re-learnt throughout life, thus allowing an opportunity for long term acquisition of graft tolerance.

The clonal theory of immune responses, and its corollary, clonal deletion as a mechanism leading to self-tolerance, were developed primarily in the context of antibody and B cells [27]. The theory was subsequently extended to T cells, and self-tolerance was proposed to result from clonal deletion in the thymus [72]. Indeed thymic tolerance

induction remains a major feature of current models of T cell function. Nevertheless, a number of features of T cell recognition distinguish it from antibody recognition, and have suggested that repertoire selection may obey a modified form of rules.

A first important difference lies in the average affinity of T cell receptor (TCR) for its antigen. At least for the subset of $\alpha\beta$ receptor carrying T cells (which is the main focus of this study), which recognise complexes of peptide presented by MHC (pMHC), this affinity is in the order of 10^{-5}M to 10^{-6}M, which is some three orders of magnitude less than that for antibody/antigen recognition [82]. In addition, only a small proportion of the TCR binding surface recognises the antigenic target peptide itself, while the rest binds to the host MHC. A consequence of these characteristics is that the individual TCRs exhibit a great deal of promiscuity and cross-reactivity: many TCRs bind to the same peptide, while many peptides can be bound by the same TCR [159, 19]. The combination of low individual affinities, and a large degree of cross-reactivity has led to the development of an elegant cooperative model of T cell recognition, the 'quorum-sensing model' [28], which proposes that functional T cell responses are the product of cooperative interactions between T cells with different receptors. The decision of whether to initiate an immune response to a particular presented peptide is made at the population level, rather being determined solely at the level of an individual T cell/antigen presenting cell ('APC') encounter.

Another fundamental distinction between T and B cells is that naive T cells require activation by antigen presented on the surface of an APC, usually a dendritic cell ('DC'). The APC provides the T cells with a high density array of MHC molecules carrying a diverse set of self and non-self peptides, but also a set of additional membrane bound and secreted signals which are necessary for productive T cell activation [134]. Dendritic cells can interact simultaneously and consecutively with many different T cells (10-20 cells at any one time, and in the order of 200-400 per hour) forming an APC/T cell cluster [103, 14, 146]. Such a cluster is an obvious candidate for the site of 'quorum-sensing', with the cluster, rather than the individual cell, acting as the unit of response. Cooperative behaviour between cells within a cluster have been documented by our group and by others [121, 35]. However, the antigen presenting activity of dendritic cells is not a static property. Dendritic cells switch from a 'resting' state, to an

'active' state, and this transition is determined to a great extent by signals from innate immunity [70]. Since resting DCs do not provide the signals necessary for naive T cell activation, they become the 'gatekeepers' of adaptive immunity, and DC activation becomes a key decision point in whether an antigenic stimulus leads to immune activation. Resting dendritic cells may not only fail to induce productive T cell activation, but may actively induce tolerance [96]. Indeed, subsets of immature dendritic cells have been shown to kill T cells in particular circumstances [166]. The concept of tolerogenic dendritic cells underlies the influential 'danger' model [95, 65], which postulates that self-tolerance results from the fact that self antigens are generally presented to T cells in the absence of innate immune responses. Thus self/non-self discrimination, at least outside the thymus, is determined as much by the dendritic cell and its interaction with innate immunity as by the T cell compartment itself.

Models for self-tolerance are still dominated by the concept of positive and negative selection operating on each individual T cell independently. The question of the mechanism for setting precise thresholds for positive or negative selection, so as to maximise response to non-self but minimise response to self, continue to be much debated [136, 47] and models have been developed that demonstrate the impact of these thresholds on the T cell response to self peptides [12, 74]. The mechanisms for establishing self-tolerance outside the thymus are also debated, although 'natural' T regulatory cells seem to play an important role [31, 6].

The very extensive literature on the induction of self-tolerance has generally been distinct from the smaller corpus of papers which deal specifically with repertoire generation. A number of models for repertoire generation have been proposed. The key experimental observations which all models must encompass are the persistently high diversity of the naive T cell pool [11], the ability for new clones to emerge and establish themselves in the repertoire [16], and the variable clone size which was an unexpected feature of the naive repertoire [125]. The majority of previous models, which often have an 'ecological' flavour, focus on clonal competition for a limited pool of presented self-antigens to drive clonal diversity and clonal size heterogeneity. Competition between T cells for access to pMHC results in stabilisation of clone sizes when all available binding sites are occupied [39] and increased diversity as those T cells that are more

different from others and therefore occupy a niche are favoured [138, 137]. In order to explain the emergence of new clones, and to prevent the development of a repertoire dominated by the clones with optimum affinities, a natural death rate of all clones is often assumed.

In the new model presented in this chapter, we combine repertoire generation and self/non-self discrimination into a single process. We incorporate aspects of cooperative behaviour (quorum sensing) into the process of naive T cell population reshaping, and explicitly model a system in which T cell receptors bind many different antigens with a range of different affinities. The model can be formalised as a linear programming (LP) optimisation problem. It shows a rapid convergence to a stable state, in which self-reactivity is maintained below a fixed threshold. The model focuses on the shaping of the T cell repertoire in the absence of immune challenge, and in this chapter we do not consider the changes to the repertoire following activation in detail. Instead we investigate the potential of the system to mount an immune response and introduce measures of the T cell population's coverage of potential non-self antigens. We show that despite the restrictions imposed by the linear constraints which ensure self-tolerance, the repertoire remains diverse, coverage is preserved and the size of individual clones becomes heterogeneous. The diversity of the constrained repertoire becomes an important factor when challenge with foreign antigens does occur, and we find that this model is able to reshape the population to retain both TCR diversity and the potential to respond to non-self more strongly than self.

3.2 Methods

3.2.1 A simple computational model

We introduce a simple computational model, and then we consider possible variations of the model and possible underlying mechanisms.

We suppose that the T cell system 'learns' in the following way to recognise self, and to react to self up to but not beyond response thresholds, which are determined by the APCs (in this study we prefer the more generic term APC, although the most important

cell type in maintenance of the naive T cell repertoire is probably the dendritic cell). Each inactive APC carries a set of self-antigens bound to MHC and is continually 'scanned' by T cells. The TCR expressed by some of these T cells recognises one of the presented pMHC complexes on the surface of the APC; T cells scan the surface of the APC, stop for a period related to the strength of interaction with pMHC and then release themselves, allowing other cells an opportunity to assay their affinity to the presented antigens [14]. In this model, we ignore any potential effects of ecological competition between T cells for pMHC binding sites in order to study the effects of the quorum sensing behaviour.

We suppose that the APC can detect the strength of the antigen specific binding between each T cell and the APC, and we further hypothesise that the APC maintains a record of the total APC/T cell binding, using some (possibly leaky) integration mechanism over a sliding time window. The APC does not need to 'know' which antigen has caused the T cell to bind, and still less which TCR clonotype the T cell expresses. The strength of signal in this model could arise from a combination of a strong affinity between pMHC and a specific TCR, or the presence of high concentrations of a particular pMHC. The model does not distinguish between these parameters but allocates an overall signal strength to each T cell/APC encounter.

We suppose that the APCs regulate the numbers of T cells in the following simple way. If the combined binding signal strength registered within a fixed time period by an APC exceeds some threshold value, then the APC sends a 'kill signal' (either actively or passively) to each T cell that is bound currently or binds subsequently [166]. These T cells, or some fraction of them, then die. Since the APC is recording the integrated signal over a sliding time window, this value will subsequently fall to below the signal threshold and the APC will then switch off the kill signal. The molecular mechanisms which could mediate such models are discussed below, but at this stage we focus on the mathematical properties of such a model.

We implement a simplified version of the model described above. The biological validity of these assumptions, and the extension of the model to more realistic but more complex scenarios is discussed later. We suppose that there are N different T cell clono-

types, with abundances at time $t = 0$ of $\mathbf{x}^0 = (x_1^0, x_2^0, \ldots, x_N^0)$. In reality the abundances would be integer counts, but in this model we treat them as positive real numbers.

We denote the binding strength between a T cell clonotype i and self-peptide MHC complex ('spMHC') k as q_{ik}. We consider a model in which each (non-activated) APC presents a particular combination (or 'profile') of spMHCs. The spMHC profile j contains an amount a_{kj} of spMHC k, and we suppose there are M such profiles that T cells may encounter. The overall binding strength of a cell of clonotype i for APC profile j is then $b_{ij} = \sum_k q_{ik} a_{kj}$. Note that when we refer to binding strength we are describing a quantity that represents the amount of signal that the APC integrates due to the T cell-APC encounter.

Each T cell may have non-zero binding strength to many spMHC complexes, and each spMHC complex may bind to many T cells: the matrix of spMHC to T cell binding strengths $Q = (q_{ik})$ is assumed to be sparse, non-negative, and with multiple positive entries in each row and column. The matrix of binding strengths of T cells to antigen profiles, $B = (b_{ij})$, therefore, is non-negative, and less sparse than Q, because each antigenic profile contains multiple spMHC complexes. B is non-negative because an APC cannot present a negative amount of antigen; that is, the a_{kj} are non-negative. Note that in this implementation we do not consider the T cell to pMHC binding strengths q_{ik}. Instead we generate the T cell to antigen profile binding strengths b_{ij} by sampling from an assumed distribution, described later.

On these assumptions, the total strength with which all T cells in the population bind to an APC with spMHC profile j is

$$r_j(\mathbf{x}) = \sum_i x_i b_{ij} \tag{3.1}$$

where $\mathbf{x}$ is the vector of clonotype abundances at time t. Writing $\mathbf{b}_j = (b_{1j}, \ldots, b_{Nj})$, we obtain:

$$r_j(\mathbf{x}) = \mathbf{b}_j \cdot \mathbf{x} \tag{3.2}$$

We set a threshold binding rate τ above which each APC will issue a kill signal to any

T cell that is bound; that is, the APC presenting self-peptide profile j issues kill signals to any T cells bound to it if $r_j(x) > \tau$. In principle, τ is a threshold that can be locally defined by the antigen presenting system: it can depend on the APC microenvironment or intrinsic antigen presenting parameters such as the MHC haplotypes. In this initial implementation, we have assumed τ is constant over all APCs. The rate at which T cell clone i is eliminated by 'kill' signals from APCs presenting self-peptide profile j is proportional to the strength of the binding interaction of each T cell with that spMHC profile j, such that:

$$\text{Kill signals for clonotype } i \text{ from APC type } j = \eta\ \phi(\mathbf{b}_j \cdot \mathbf{x} - \tau_j) b_{ij} x_i \tag{3.3}$$

where η is a rate parameter, $\phi(\mathbf{b}_j \cdot \mathbf{x} - \tau_j)$ is the fraction of all T cells binding to APC j that receive a kill signal, and x_i is the abundance of T cells of type i. Our hypothesis is that kill-signals are only issued when the rate of binding to APCs is greater than τ; this hypothesis is expressed in terms of the function $\phi(z)$, which is some non-decreasing function such that $0 \leq \phi(z) \leq 1$ for all real z. $\phi(z)$ should be small or zero for $z < 0$, and we suppose that $\phi(z)$ rises towards 1 rapidly for $z \geq 0$. The simplest choice for ϕ would be the Heaviside function $H(z) = 1$ if $z \geq 0$ and $H(z) = 0$ if $z < 0$; a more biologically realistic function would be continuous and differentiable, such as the logistic function $\phi(z) = \frac{1}{1+\exp(-\alpha z)}$, for some suitable scale parameter α. The implementation captured in equation 3.3 further assumes each APC, and hence each spMHC profile j occurs once, but the model is easily extended to incorporate variable APC numbers for each antigen profile.

So far, the model only has a mechanism for killing T cells: there must also be a method for T cells to multiply. Although it is clear that naive cells must see self-antigens in order to survive, the quantitative relationship between antigen binding strength and proliferation in the context of T cell homeostatic proliferation remains unclear. Here we adopt the simplest assumption, namely that all T cells spontaneously divide at some rate ν, although a model relating ν to binding strength could also be implemented.

Using these assumptions, we obtain that for each clonotype i:

$$\dot{x}_i = \nu x_i - \eta \sum_j \phi(\mathbf{b}_j \cdot \mathbf{x} - \tau_j) b_{ij} x_i \tag{3.4}$$

so that

$$\frac{\dot{x}_i}{x_i} = \nu - \eta \sum_j \phi(\mathbf{b}_j \cdot \mathbf{x} - \tau_j) b_{ij} \tag{3.5}$$

We can demonstrate rather simply that the optimisation will indeed always converge. For a suitable choice of ϕ the right hand side can be written as the gradient of a convex function of $\mathbf{x}$. For a function with a second derivative, it is convex on an interval if its second derivative is non-negative in that interval, is strongly convex if the second derivative is positive, and is concave if the second derivative is non-positive. Observe that:

$$\Phi(u) = \int_{-\infty}^{u} \phi(z) dz \tag{3.6}$$

exists for plausible choices of ϕ, and is convex and differentiable provided that $\phi(z)$ is non-decreasing and continuous. Then define:

$$f_j(\mathbf{x}) = \Phi(\mathbf{b}_j \cdot \mathbf{x} - \tau_j) \tag{3.7}$$

Each f_j is convex in $\mathbf{x}$, and note that:

$$\frac{\partial f_j(\mathbf{x})}{\partial x_i} = \phi(\mathbf{b}_j \cdot \mathbf{x} - \tau_j) b_{ij} \tag{3.8}$$

Now define:

$$F(\mathbf{x}) = -\nu \sum_i x_i + \eta \sum_j f_j(\mathbf{x}) \tag{3.9}$$

which is a sum of convex differentiable functions. The scalar function $F(\mathbf{x})$ is constructed so that

$$\frac{\partial F(\mathbf{x})}{\partial x_i} = -\nu + \eta \sum_j \phi(\mathbf{b}_j \cdot \mathbf{x} - \tau_j) b_{ij} = -\frac{\dot{x}_i}{x_i}$$

so that the rate of change of $\mathbf{x}$ is expressed as

$$\dot{x}_i = -x_i \frac{\partial F(\mathbf{x})}{\partial x_i}$$

We can now write the rate of change of $F(\mathbf{x})$ as:

$$\frac{dF(\mathbf{x})}{dt} = \dot{\mathbf{x}} \cdot \nabla F(\mathbf{x}) \tag{3.10}$$

$$= -\sum_i x_i \left(\frac{\partial F(\mathbf{x})}{\partial x_i} \right)^2 \tag{3.11}$$

$$\leq 0 \text{ since all } x_i \text{ are positive} \tag{3.12}$$

F is convex and differentiable, because it is the sum of convex, differentiable functions, and F therefore has a unique minimum in the region of interest, which is the non-negative quadrant. At this minimum, all constraints $\mathbf{b}_j \cdot \mathbf{x} \leq \tau_j$ will be approximately satisfied, provided that the growth rate ν is small compared to the 'kill rates' from the APCs.

From equation 3.11, we know that the value of F, which includes a sum of measures of constraint violation, must decrease over time. However, it says little about the rate of convergence towards the minimum of F. In Appendix C, we present a stronger analysis of the convergence of the process of equation 3.4, by identifying it with a version of the multiplicative weight updating algorithms surveyed by [10]. This analysis establishes regret bounds for such updates on a possibly time-varying set of constraints. We note that equation 3.4 could be solved by standard differential equation methods, provided the rate of killing (and the rate of proliferation) remain constant. Under these conditions, the iterations become equivalent to a fixed time step, which can be allowed to decrease to the continuous case. However, we prefer to use the iterative algorithm we describe because the discrete time steps are readily interpretable in terms of cellular events (e.g. T cell/APC interactions) and because the regret bounds it establishes are robust to variations in rate. The model therefore leaves open the possibility of introducing time-dependent and tissue-dependent variations in rates in future extensions of the basic model.

The implementation described here gives rise to a series of constraints on T cell abundances, which are captured by a series of linear inequalities as outlined above. An iterative method to solve this linear programming problem is set out below, and can be given a feasible biological interpretation. Note that the proliferation rate ν is set so that in the absence of any 'kill signals' the T cell population would double in one unit of time, and the rate η of T cell killing is set relative to this.

1. Calculate the immune response to each profile, $r_j \leftarrow \sum_i x_i b_{ij}$ for all j.

2. Determine for which self-profiles the response threshold has been violated, $v_j \leftarrow [r_j > \tau]$ for all j.

3. Adjust the T cell clonotype abundances, $x_i \leftarrow x_i \left(1 + \nu\ \delta t - \eta\ \delta t \sum_j b_{ij} v_j\right)$

The multiplicative update analysis discussed in Appendix C provides strong guarantees for time-varying constraints, corresponding to the case where APCs present varying combinations of antigens over time.

3.2.2 Assessing the potential for an immune response

In order to investigate the potential of the reshaped T cell population to mount an immune response to previously unencountered antigens, we create a set of new independently generated antigenic profiles which were not part of the set on which the T cell population has been trained. We refer to these as 'non-self profiles'. The binding strength of each existing TCR for each new profile is selected independently of its given affinities for all the self profiles, although the value is selected from the same probability distribution. We use these non-self profiles to test whether under our assumptions the T cell repertoire will achieve the dual objectives of maintaining self-tolerance, while at the same time maintaining as broad and strong a repertoire for non-self as possible.

Note that we do not model an immune response to these new profiles here. If the APC remains in a tolerogenic state, the introduction of new non-self profiles will typically violate the constraints, but this will result in additional T cell killing and the system will gradually readjust to remain within the immune activation threshold. We envisage that if the APC were switched to an immunogenic state (for example by exposure to

innate immune danger signals) then crossing the threshold would result in activation of all APC bound T cells, resulting in an effector immune response.

We measure the ability of the T cell population to respond to a non-self profile as the total potential T cell response, calculated as $r_j(\mathbf{x}) = \sum_i x_i c_{ij}$ for non-self profile j, where $C = (c_{ij})$ is the matrix of binding strengths between T cell clonotypes and non-self profiles. It is important to note that we are not simulating the behaviour of the T cell population on immune challenge here, but assessing the potential of the reshaped repertoire to respond to previously unencountered profiles. In order to measure the 'success' of the reshaped repertoire we can consider its coverage of the potential non-self antigen space. The first coverage measure we use in this study is the ratio of the mean total repsonse against non-self profiles to the mean total response against self profiles: coverage $= \frac{\overline{r_{ns}}}{\overline{r_s}}$ for self profiles s and nonself profiles ns. Alternatively, we also measure the coverage as the proportion of non-self profiles that give a potential T cell response greater than the average response to self profiles, i.e. $|\{ns : r_{ns} > \overline{r_s}\}|$ for nonself profiles ns expressed as a fraction of the total number of nonself profiles modelled.

3.3 Results

3.3.1 Clone size adjustment algorithm reaches a solution of the repertoire constraints: violations are resolved rapidly and repertoire is optimised slowly

We first simulate a very simplified repertoire to allow us to visualise the action of the update algorithm. We start with two T cell clonotypes and three spMHC profiles. The clonotypes have binding strengths for each of the profiles as detailed in Figure 3.1g. In this simulation each profile is given the same total response threshold ($\tau = 1$), above which there will be harmful autoimmunity. The other parameters of the update algorithm are set out in the legend of Figure 3.1.

The self-response thresholds for each profile and the binding strengths between clonotypes and spMHC profiles (Figure 3.1g) give constraints on allowable repertoires. If x_i

is the abundance of clonotype i, to avoid autoimmunity we require that:

$$1.2x_1 + 1.0x_2 \leq 1$$

$$1.5x_1 + 0.5x_2 \leq 1$$

$$0.6x_1 + 1.4x_2 \leq 1$$

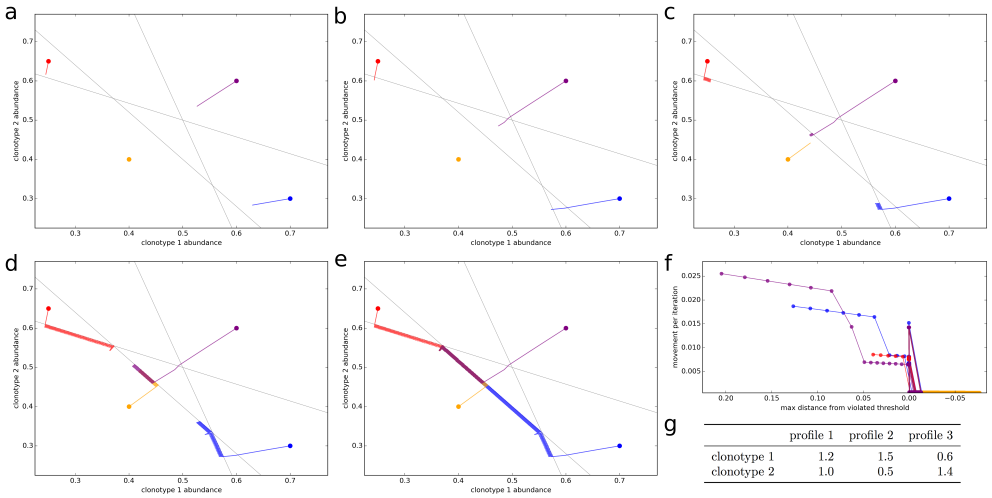

Figure 3.1: Optimisation of the T cell population to avoid autoimmunity while maximising T cell numbers in a simplified system.
A simplified repertoire containing two clonotypes and three spMHC profiles. The update algorithm is initiated with different initial clonotype abundances each represented by a different colour. The coloured lines track the changes in clonotype abundance over iterations of the update algorithm.
(a)-(e) The clone abundances after (a) 5, (b) 10, (c) 100, (d) 1000 or (e) 10,000 iterations of the update algorithm. The grey lines indicate the constraint that total T cell response should be less than the threshold for each of the spMHC profiles.
(f) For each of the starting repertoire configurations, the relationship between the Euclidean distance moved by the repertoire configuration in an iteration of the update algorithm and the distance from the furthest violated threshold, or if there are no violations the distance to the nearest threshold.
(g) The affinities between clonotypes and spMHC profiles.
Other model parameters for all panels are: $\tau = 1$, the self-response threshold for each spMHC profile, $\nu = \ln 2 \ \delta t^{-1}$, the growth rate and $\eta = 0.01001 \ \delta t^{-1}$, the learning rate.

We repeatedly simulate the update algorithm with different starting repertoire configurations. Each starting configuration is represented as a colour in Figure 3.1. The panels in this figure show a time course of the update algorithm working on each initial repertoire configuration.

We see that if the initial repertoire configuration violates one or more of the response

constraints, the update algorithm very quickly shapes the repertoire (by adjusting clonotype abundances) to a point where there is no autoimmunity (Figure 3.1b, 10 iterations). In contrast, when no threshold is violated by the starting configuration of the repertoire (yellow path in Figure 3.1) the repertoire does not move very far from the initial configuration in the first cycles.

Once the repertoire has been moved to a configuration where all constraints are satisfied, the update algorithm continues to allow each clonotype to become as abundant as possible while remaining inside the 'feasible region' (Figure 3.1c-e). For this arrangement of affinities, the 'optimum' repertoire in terms of having the highest total abundance while avoiding autoimmunity is at a single vertex of the feasible region, and we can see that the update algorithm moves each of the initial repertoires slowly towards this point.

The speed which which the clonotype abundances are adjusted is dependent on the severity of the violation of the thresholds, as the update rule is designed to do through the negative learning rate η. This can be quantified by considering the Euclidean distance moved by the repertoire configuration in a timestep as a function of the Euclidean distance by which the current configuration violates a threshold (Figure 3.1f). There is a strong positive relationship between the severity of the violation and the speed with which the update algorithm adjusts the clonotype abundances.

3.3.2 Positive selection of clonotypes based on self-profile binding strength is required for successful immune tolerance

We next simulated the update algorithm with a larger number of T cell clonotypes and spMHC profiles (Figure 3.2). For each clonotype-profile pair, the binding strength (b_{ij} for clonotype i and profile j) is set to zero with probability $1-\gamma$. If the binding strength is not set to zero it is selected at random from a left-censored normal distribution. For simplicity we set the response threshold τ to be equal to 1 for all self profiles.

We run the update algorithm and record the abundance of each clonotype at each iteration. Note that under our constant growth rate assumption, iterations can be thought of as directly equivalent to T cell generations. We set the growth rate ν such that one

unit time is equal to one T cell generation, giving one T cell generation in approximately 1387 iterations. The total T cell response to a spMHC profile can be calculated as the sum of abundance $\times$ binding strength for each T cell clonotype ($r_j = \sum_i x_i b_{ij}$ for spMHC profile j). We can then define successful immune tolerance as the reshaping of the T cell population into one where the total T cell response to any spMHC profile (r_j for self-profile j) is below the threhsold τ. The mean total response to spMHC profiles over time (Figure 3.2a, solid line) is initially well controlled at the allowed threshold. However, after approximately 10,000 iterations of the update algorithm the control of the response breaks down and there is an increased average response to self, above the allowable threshold.

We noted that those clonotypes that are highly abundant after running our simulation for 30,000 cycles of the update algorithm have low maximum binding strength to spMHC profiles. We can see the reason for breakdown of control of self response if we consider a clonotype of abundance 1 that has zero binding strength for all self profiles except one, for which it has binding strength b. Then after one iteration of the update algorithm, the clonotype will have abundance $(1+\nu\ \delta t)$ or $(1+\nu\ \delta t - \eta\ \delta t\ b)$ depending on whether the total T cell response to the profile for which it has non-zero binding strength is below the allowable self-response threshold τ or not. In order to avoid uncontrolled growth of the clonotype, we would require that $(1+\nu\ \delta t - \eta\ \delta t\ b) < 1$, which is equivalent to requiring that $b > \nu/\eta$. Therefore we suggest that the inability of the update algorithm to control average self response is due to the presence of clonotypes for which the maximum binding strength to any of the self profiles is below ν/η. This indicates the requirement for some form of positive selection.

In its simplest form, positive selection would take the form of a function which deletes all clones whose maximum binding strength for any self profile is below ν/η. A more realistic function could make the growth rate in any one cycle depend on the average binding strength to self profiles or to the maximum binding strength to a randomly selected sample of 'encountered' self profiles. In the following work we implement the simplest form of the affinity-dependent selection, by eliminating all clonotypes with maximum binding strength to self profiles below ν/η before the update algorithm begins.

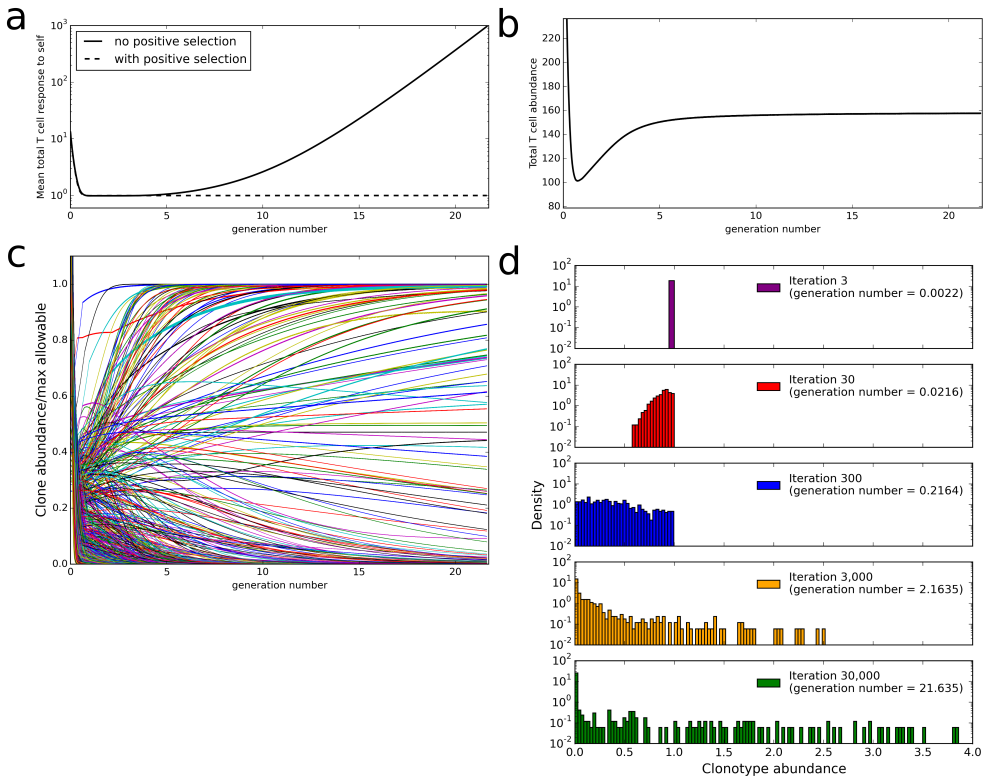

Figure 3.2: Evolution of the repertoire under the constraints of dendritic cell dependent T cell deletion.

(a) mean total T cell response to spMHC profiles over time as clone sizes are updated according to the basic immune tolerance learning algorithm with or without positive selection implemented. The total T cell response to a profile is calculated as the sum of (abundance $\times$ binding strength) for each clonotype.

(b) the total T cell abundance (being the sum of the abundances of clonotypes present at a particular time) in the simulation after T cell positive selection is implemented

(c) the abundance of each T cell clonotype over time after T cell positive selection is implemented. The abundance is expressed as a fraction of the maximum abundance the clonotype could reach without violating any self-response thresholds if it was the only clonotype in the population.

(d) the clone size distribution after the indicated amount of time after T cell positive selection is implemented.

Model parameters for all panels are: self-response threshold $\tau = 1$, growth rate $\nu = \ln 2 \ \delta t^{-1}$, learning rate $\eta = 0.002001 \ \delta t^{-1}$, number of spMHC profiles $M = 100$, number of T cell clonotypes $N = 1000$, proportion of non-zero affinities $\gamma = 0.01$.

We implement this positive selection of clonotypes and re-run the simulation with the same parameters (detailed in Figure 3.2 legend). The total T cell response to self profiles is now tightly controlled at the allowable threshold (Figure 3.2a, dashed line). It is however possible under this model that if a clonotype escapes positive selection it slowly increases in size indefinitely.

3.3.3 Total population size homeostasis but increased clonotype abundance heterogeneity as a function of time

Naive TCR repertoires are made up of clonotypes with a broad range of abundances. We therefore examined the abundance distribution produced by the model presented in this chapter. Since our implementation of the model uses continuous rather than discrete abundances, abundances never reach zero but become arbitrarily small. In order to consider the abundance distribution, we therefore set a lower threshold below which a clone is considered to be deleted. In this work we consider a clonotype to be completely absent when its abundance falls below a threshold defined by $N/10^8$ where N is the number of clonotypes in the simulation. This threshold was chosen based on consideration of a mouse immune system which has in the order of 10^8 T cells in total. If N different clonotypes of equal abundance are present in this repertoire, each clonotype could be considered to have a starting abundance of $10^8/N$. Hence if a clone contracts by a factor of $> 10^8/N$, its abundance would fall below 1 and hence the clonotype can be considered as eliminated. Since the abundance of each clonotype at the start of the model is arbitrarily initiated at a value of one, this is equivalent to defining a clone with an abundance of lower than $N/10^8$ as deleted.

We first considered the total size of the T cell compartment as a function of time. At every timepoint during the simulation we can calculate the total size of the repertoire as the sum of the clonotype abundances that are above the 'presence' threshold of $N/10^8$ (Figure 3.2b). We see that this initially contracts as self-response constraint violations are resolved, but then expands (driven by the positive learning rate increasing the abundance of each clonotype when constraints are not violated) until a stable level is reached where growth and negative selection are balanced. If all other parameters of the model are fixed, the eventual total size of the T cell compartment at homeostasis is strongly

correlated to the number of clonotypes present in the repertoire at the beginning of the simulation.

We then consider the abundances of individual clonotypes. The maximum allowable size, m_i for a clonotype in the model can be defined as the self-response threshold divided by the maximum binding strength the clonotype has for any self profile, i.e.

$$m_i = \frac{\tau}{\max_j b_{ij}}$$

is the maximum allowable size for clonotype i. For each of the clonotypes in the simulation, we consider its abundance, expressed as a proportion of the maximum allowable abundance m_i for that clonotype, across time (Figure 3.2c). Some clonotypes are present close to their maximum allowable size m_i, presumably due to lack of cross-reactivity with other profiles or other clonotypes, while some clonotypes are quickly removed from the repertoire. It it interesting to note that while the total T cell abundance stabilises rapidly (Figure 3.2b) the individual clonotype sizes continue to be adjusted even in later stages of the simulation. The clone size distribution (Figure 3.2d) spreads to include smaller clonotypes during the initial part of the simulation, then starts to include larger clonotypes as well in later iterations. At the end of our simulation there is a large spread of clone sizes in which large and small clones co-exist, as observed experimentally, rather than a repertoire completely dominated by a few large clonotypes.

3.3.4 Increased number of T cell clonotypes provides greater repertoire coverage

A successful T cell population needs to be able to control immune response to self but at the same time must provide broad coverage against a range of unknown non-self antigens that the individual might encounter. The mean total potential T cell response to self and non-self profiles ($\pm$ standard deviation) across iterations is shown for one set of simulation paramters in Figure 3.3a.

This shows that the response to self is well controlled at the allowed threshold τ. In

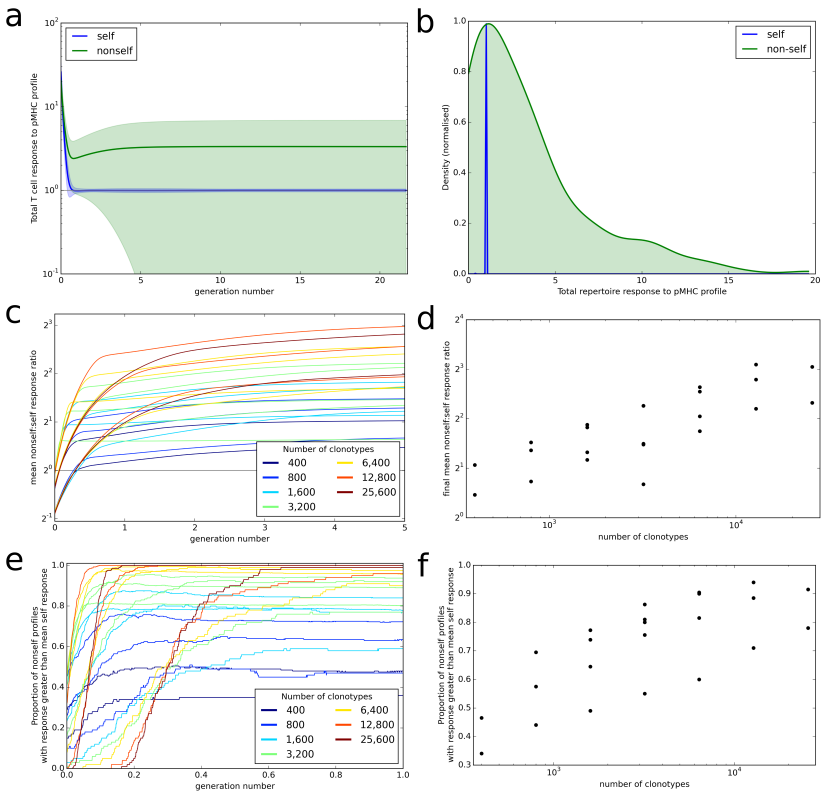

Figure 3.3: Broad coverage to non-self is maintained during the development of a self-tolerance repertoire.

(a) The mean ($\pm$ standard deviation) total T cell response to self (blue) or non-self (green) pMHC profiles over time with $N = 2000$ and $M = 200$.

(b) After 30,000 iterations of the update algorithm with parameters as in (a), the distribution of total T cell response to self (blue) and non-self (green) pMHC profiles.

(c) The ability of the repertoire to successfully mount an immune response to non-self pMHC profiles, measured as the average total response to a non-self profile divided by the average total response to a self profile, over time. The number of T cell clonotypes in a simulation is indicated by colour, with the number of self profiles simulated ranging between 100 and 800.

(d) The relationship between number of T cell clonotypes and the average total response to a non-self profile divided by the average total response to a self profile after 30,000 iterations of the update algorithm.

(e) The proportion of non-self profiles that have a total T cell response greater than the mean response towards self profiles over time. The number of T cell clonotypes is indicated by colour.

(f) The relationship between the number of T cell clonotypes and the proportion of non-self profiles having a stronger total T cell response than the mean response to self profiles after 30,000 cycles of the update algorithm.

Other model parameters for all panels are: self-response threshold $\tau = 1$, growth rate $\nu = \ln 2 \ \delta t^{-1}$, learning rate $\eta = 0.002001 \ \delta t^{-1}$ and proportion of non-zero affinities $\gamma = 0.01$.

contrast, the average response to nonself pMHC profiles becomes higher as the model shapes the repertoire. However, the nonself responses become very heterogeneous. After 30,000 iterations the response to all self profiles is at or near the allowed threshold while the majority of nonself profiles are capable of more T cell binding, and therefore a larger potential T cell response (Figure 3.3b). However, there are also a number of nonself profiles that create a lower response than that to self profiles. These presumably represent 'holes' in the repertoire coverage.

We assess the ability of the reshaped repertoire to cover the potential nonself antigen pool via the two coverage measures described earlier. We ran the update algorithm a number of times with the number of T cell clonotypes (N) ranging between 400 and 25,600 and number of spMHC profiles (M) ranging between 100 and 1600 (only running combinations where $M < N$). Other parameters of the simulation are detailed in Figure 3.3 legend. We considered the evolution of the ratio of mean self potential response to mean nonself potential response across cycles of the algorithm (Figure 3.3c) and see that this increases until it is above 1 (indicating higher potential response to nonself than self profiles) for all parameter sets. The repertoire coverage, using this measure, depends on the total number of T cell clonotypes in the repertoire at the start of the algorithm (Figure 3.3d).

The proportion of nonself profiles that the T cell population has the potential to respond to more strongly than it does to self profiles is initially low but is increased as the update algorithm shapes the repertoire (Figure 3.3e). The success of the repertoire under this measure is again strongly correlated to the number of clonotypes (Figure 3.3f).

3.3.5 Clonotype diversity and spMHC profile cross-reactivity are preserved by the update algorithm

We have demonstrated that the model described in this study produces a TCR repertoire that respects self-response thresholds, but violates the thresholds when exposed to non-self antigen profiles. It has been observed that the TCR repertoire in an individual remains diverse (many different clonotypes are present, with cross-reactivity between clonotypes and profiles) until old-age, when a few dominant clonotypes appear

[25]. We explored whether our selection model can retain diversity in the repertoire or whether the multiple linear constraints favour a sparse solution with few surviving clonotypes.

We first consider the proportion of starting clonotypes surviving (i.e. with an abundance greater than the lower limit defined earlier) as a function of time. The proportion of clonotypes present in the repertoire falls rapidly in the initial stages of repertoire reshaping and then stabilises (Figure 3.4a, blue). The proportion of the initial clonotypes that remain after 30,000 cycles of the update algorithm is inversely correlated to the the number of clonotypes in the simulation (Figure 3.4b, blue).

A key parameter of the adaptive immune system is the amount of information it can encode. The information content encoded in the repertoire (which depends on a combination of the number of different T cell clones and their relative size) can be captured by the Shannon Information (SI) coefficient, which is the log of the true diversity of order 1 [130]. The SI coefficient of the repertoire initially decreases rapidly before stabilising (Figure 3.4a, red). However, there is only a weak (and not statistically significant) correlation between the Shannon Information coefficient and the number of clonotypes in the simulation (Figure 3.4b, red).

Cross-reactivity, such that multiple TCRs can recognise the same pMHC profile and multiple pMHC profiles can be recognised by the same TCR, is a well-recognised feature of the T cell repertoire [159, 19]. To investigate the evolution of cross-reactivity in our model, we measure the number (or proportion) of clonotypes which have non-zero binding strength for a single pMHC profile (i.e. $|\{i : b_{ij} > 0\}|$ for each profile j). The mean proportional cross-reactivity against self profiles decreases initially then begins to stabilise, while the mean cross-reactivity against nonself profiles is maintained (Figure 3.4c).

After running the simulation for 30,000 iterations of the update algorithm, the distributions of cross-reactivity against self and nonself profiles are clearly different (Figure 3.4d). The majority of nonself profiles are recognised by more TCR clonotypes than self profiles are, and the ratio of self:nonself cross reactivity is not significantly correlated to the size of the simulation (Figures 3.4e and f).

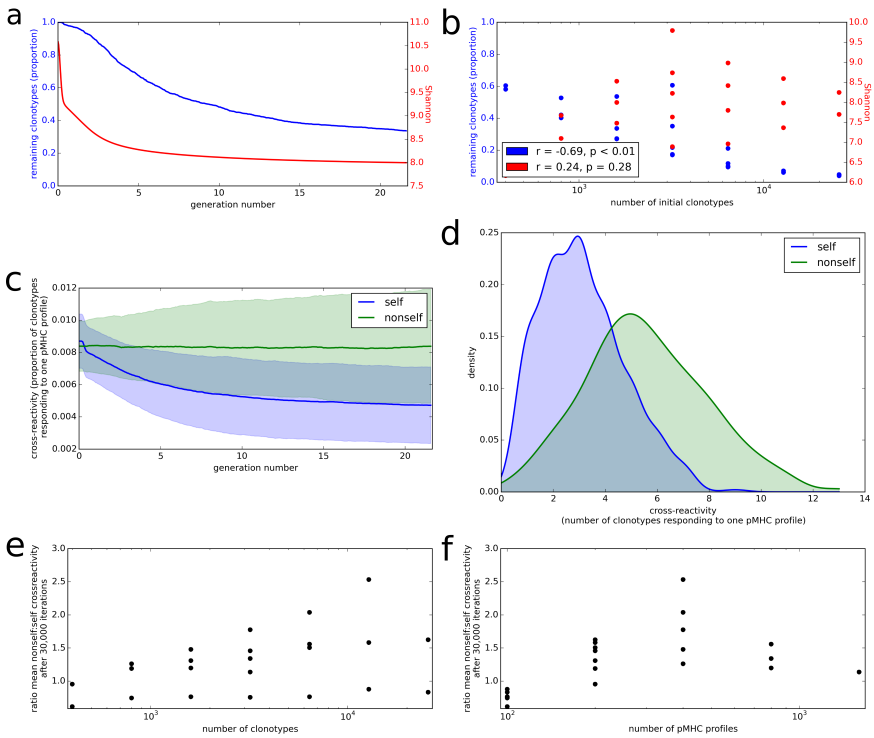

Figure 3.4: Clonotype diversity and pMHC profile cross-reactivity are preserved by the update algorithm.

(a) Blue: The proportion of clonotypes (after positive selection) that are present over time during simulation of the update algorithm. Red: The Shannon entropy of the repertoire over time. Simulation implemented with $N = 1600$ and $M = 400$.

(b) Relationship between number of clonotypes in the simulation and proportion of clonotypes remaining (blue) or Shannon entropy of the repertoire (red) after 30,000 iterations of the update algorithm. Simulation implemented with values of N between 400 and 25,600 and M beween 100 and 800, with $M < N$.

(c) Cross-reactivity of T cell clonotypes against self (blue) and nonself (green) pMHC profiles over time, run with $N = 3200$ and $M = 400$. Cross-reactivity is measured as the proportion of present clonotypes that have non-zero binding strength for a given profile. Data shown is mean cross-reactivity across all profiles $\pm$ standard deviation.

(d) Distribution of cross-reactivity across all self (blue) and nonself (green) pMHC profiles after 30,000 iterations of the update algorithm with $N = 3200$ and $M = 400$. Cross-reactivity is measured as the absolute number of present clonotypes that have non-zero binding strength to a profile.

(e) Relationship between the number of clonotypes present at the start of the update algorithm and the ratio of the mean cross-reactivity against nonself profiles to the mean cross-reactivity against self profiles after 30,000 cycles of the update algorithm.

(f) Relationship between the number of self profiles in the update algorithm and the ratio of the mean cross-reactivity against nonself profiles to the mean cross-reactivity against self profiles after 30,000 cycles of the update algorithm.

Other model parameters for all panels are: self-response threshold $\tau = 1$, growth rate $\nu = \ln 2 \ \delta t^{-1}$, learning rate $\eta = 0.002001 \ \delta t^{-1}$ and proportion of non-zero affinities $\gamma = 0.01$.

3.3.6 New clonotypes can establish themselves in a stable repertoire

The TCR repertoire is constantly being updated by the introduction of new T cells from the thymus, and new clonotypes can establish themselves despite competition from the existing clonotypes. We explored whether the update algorithm of our model would allow introduction of new clonotypes. We ran the update algorithm for 30,000 iterations to produce a self tolerant and stable repertoire and then selected 10 of the clonotypes present at random. We created 10 new duplicate clonotypes, with identical spMHC profile binding strength values as the selected clonotypes, and introduced them into the repertoire at an abundance equal to the average abundance of the exisiting clonotypes. We then tracked both the original 10 clones, and their duplicates, for further iterations of the simulation.

The total T cell abundance increases transiently as new clonotypes are introduced but quickly returns to a stable level (Figure 3.5a). On introduction of the new duplicate clonotypes the abundances of the original 10 clonotypes all fall in order to satisfy the self-response constraints (Figure 3.5b). Clonotypes with matched self-binding strength profiles are seen to tend towards the same abundance over the additional iterations of the model (Figure 3.5c and d). Although the abundances of the new clonotypes do not reach equality with the original clonotypes, the introduced clonotypes only disappear in cases where the original clonotypes are also deleted. The introduced clonotypes are able to remain in the repertoire even when they are introduced at a lower abundance than an already established clonotype with the same self-response profile.

To further test the ability of new clonotypes to establish themselves in a self-tolerant repertoire, and to investigate whether we can predict whether a clonotype will be incorporated into the repertoire or removed, we perform the same experiment except that the introduced clonotypes have similar but not identical affinity profiles to the existing clonotypes.

We create affinity profiles for new clonotypes by selecting an existing clonotype and adding ‘noise’ to its affinity profile as follows. If the existing clonotype has affinity b_j

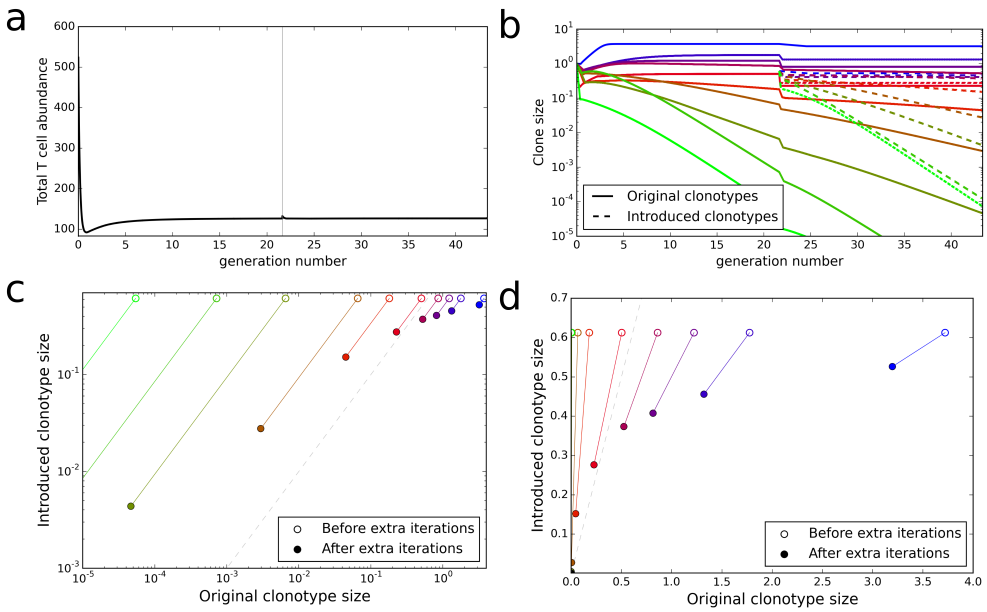

Figure 3.5: New clonotypes can establish themselves in a stable repertoire.
(a) Total T cell abundance over time. 10 new clonotypes, each with self profile binding strength vector matching an existing clonotype, are introduced (at the average clonotype abundance) after 30,000 iterations of the update algorithm.
(b) The clonal abundance of 10 selected clonotypes (solid lines) over time. After 30,000 iterations of the update algorithm, 10 additional clonotypes are introduced (dashed lines), each with a self profile binding strengths equal to one of the original 10 clonotypes. Colours represent binding strength profiles (selected clonotypes).
(c) and (d) For each of the selected original clonotypes and the binding strength-matched introduced clonotypes, the relationship between the original and match clone abundance when the new clonotype is introduced (open circles) and after running the simulation for an additional 30,000 iterations of the update algorithm (solid circles). The dashed grey line represents identical abundance of original and introduced clonotypes.
Model parameters used for all panels are: self-response threshold $\tau = 1$, growth rate $\nu = \ln 2\ \delta t^{-1}$, learning rate $\eta = 0.002001\ \delta t^{-1}$, proportion of non-zero affinities $\gamma = 0.01$, number initial clonotypes $N = 1000$, number self profiles $M = 100$.

to self-peptide profile j then the affinity of the new clonotype for self-peptide profile j is chosen from $N(b_j, fb_j)$ where f is a factor which defines the amount of 'noise' being introduced. Ten clonotypes present after 30,000 iterations of the update algorithm are selected, at a range of abundances, and new clonotypes are introduced each with the affinity profile of one of the selected clonotypes as a template. Different amounts of 'noise' are considered, using $f = 0, \frac{1}{8}, \frac{1}{4}$, or $\frac{1}{2}$. The new clonotypes are introduced into the repertoire at the current average clone abundance and the simulation is allowed to run for 30,000 further iterations.

For each noise factor, the total T cell population quickly returned to homeostasis and

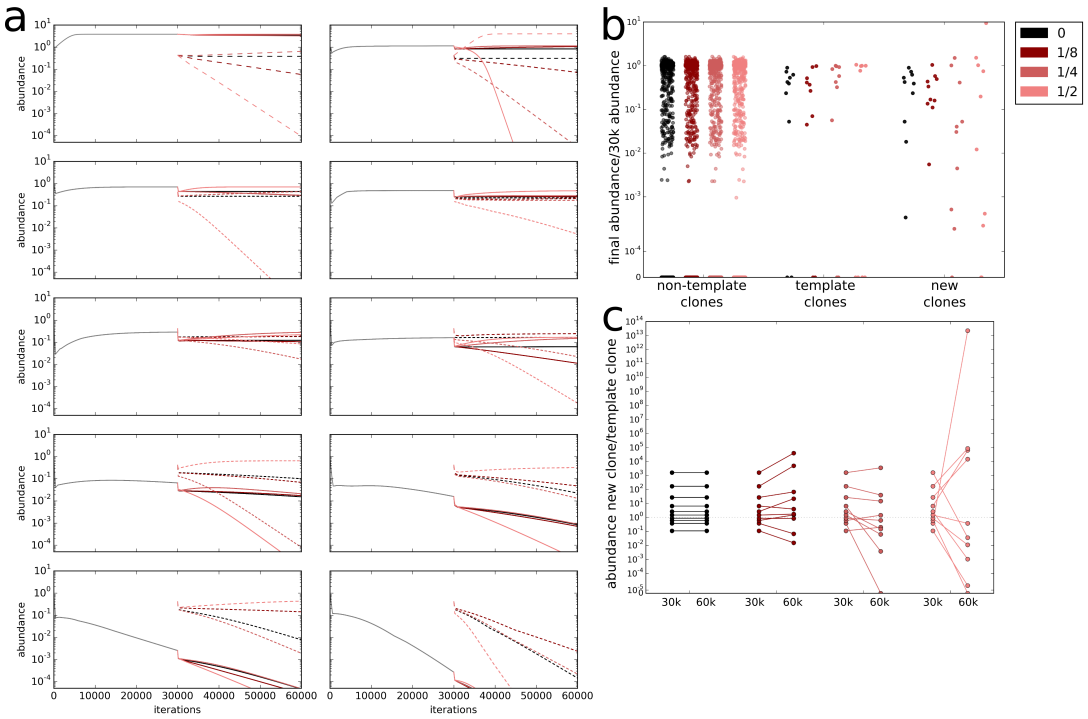

Figure 3.6: New clonotypes introduced with 'similar' affinity profiles to existing clonotypes demonstrate diverse behaviour

The amount of noise applied to template clonotype affinity profile to create the profile for the new clonotype is indicated by colour. Elements of the new affinity profile are selected from a normal distribution with mean equal to the equivalent element of the template affinity profile and standard deviation equal to the 'noise factor' (given by colour) multiplied by the element of the template profile.

(a) For each of ten template clonotypes (solid lines), the abundance of the template clonotype over time, and the abundance of the introduced clonotypes (dashed lines) over time after introduction at 30,000 interations.

(b) Clone abundance at 60,000 iterations over clone abundance at 30,000 for clones that were: (i) not used as templates (left hand group), (ii) used as templates (middle group) and (ii) newly introduced (right hand group).

(c) For each template and new clonotype, the abundance of the new clonotype divided by the abundance of the template clonotype at 30,000 iterations (left hand of each pair) and 60,000 iterations (right).

All simulations run with parameters: self-response threshold $\tau = 1$, growth rate $\nu = \log 2 \delta t^{-1}$, learning rate $\eta = 0.0020001 \delta t^{-1}$, proportion of non-zero affinities $\gamma = 0.01$, number initial clonotypes $N = 5000$, number self-profiles $M = 500$.

immune tolerance was restored, as was previously seen for $f = 0$ (Figure 3.5). The abundance of each original and introduced clone over iterations (Figure 3.6a) shows that the template clones that are smallest originally are removed from the repertoire on introduction of the new clonotypes, as are the new clonotypes with affinity profile derived from these smallest clones, regardless of the amount of noise applied to the copied self-affinity distributions. This pattern was repeated in two other runs of the simulation, not shown.

The abundance of the majority of the existing clonotypes which were not used as templates for introduced clonotypes stayed approximately constant between 30,000 and 60,000 iterations (Figure 3.6b, left) regardless of the amount of noise introduced, while some were completely removed from the repertoire and others decreased in size. A similar pattern was observed in the abundance of existing clonotypes that were used as templates for the new clonotypes (Figure 3.6b, middle), although more decreased in size or were removed than remained the same size. The reduction in abundance of these clones appears to be correlated to the amount of noise. The introduced clonotypes are the only ones which are able to increase in abundance between 30,000 and 60,000 iterations, although most are also reduced (Figure 3.6b, right). Increase in clone abundance only appears to be possible when there is non-zero noise.

For each template clone and the new clonotype based on its affinity profile, comparing their abundances at 30,000 and 60,000 iterations (Figure 3.6c) demonstrates that, when the template is copied with no noise, the proportional clone sizes, in relation to each other, remain constant. However, including noise when the template affinity distribution is being copied results in relative clone sizes changing between 30,000 and 60,000 iterations. Some introduced clones increase in size in comparison to their templates, while others decrease. We were interested to determine what were the particular properties of the new clones which allowed them to compete and expand within the existing stable repertoire. In order to investigate this question the abundance of the new clones at 30,000 iterations after introduction was compared to a number of characteristics of the repertoire.

When affinity profiles are created from templates with the addition of noise, the final

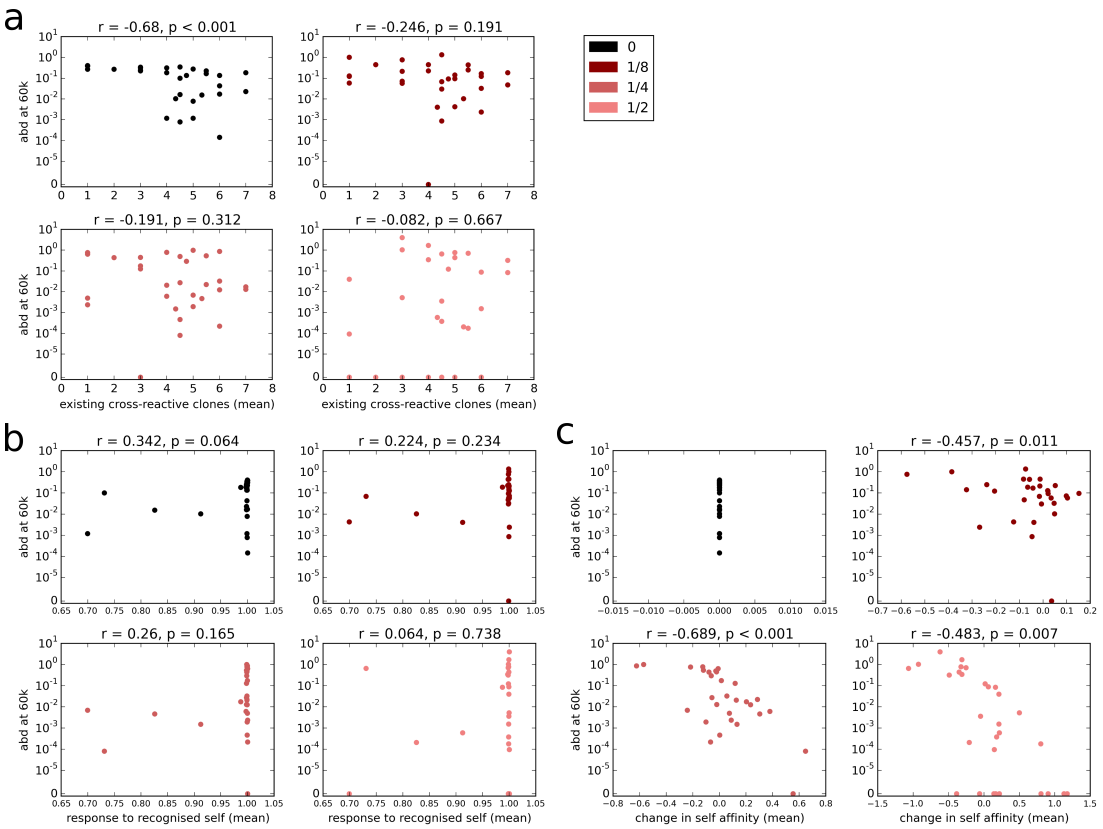

Figure 3.7: Ability of a clone to establish itself in the repertoire depends on change in affinity profile from template
Relationship between abundance of new clonotypes 30,000 iterations after introduction and (a) the mean number of other clonotypes with affinity for the self-peptide profiles that the new clonotype has non-zero affinity for, (b) the total population wide affinity to those self-peptide profiles and (c) the mean change in affinity from the template clone to the new clone.
All simulations run with parameters: self-response threshold $\tau = 1$, growth rate $\nu = \log 2\delta t^{-1}$, learning rate $\eta = 0.0020001\delta t^{-1}$, proportion of non-zero affinities $\gamma = 0.01$, number initial clonotypes $N = 5000$, number self-profiles $M = 500$.

abundance of a new clone is not correlated to the mean number of other clones in the established repertoire which have affinity to the same self-peptide profiles as the introduced clone (Figure 3.7a), nor to the mean total potential response (affinity $\times$ abundance) to those self-peptide profiles that exists in the repertoire before introduction of the new clones (Figure 3.7b).

However, when the final abundance of a clone is compared to the mean change in affinity from the template clonotype, there is a statistically significant negative correlation (Figure 3.7c). The clones that are able to most effectively establish themselves in the repertoire are those that on average have a lower affinity to self peptide profiles than

their template molecules. This suggests that in order to out-compete an already established clone a new clonotype which recognises the same self-peptides will need to do so with lower affinity.

3.4 Discussion

We have outlined a simple computational model by which the peripheral T cell repertoire in an individual can be continually adjusted in order to optimise the chance of a successful response to unknown pathogens while minimising the amount of dangerous T cell response to self. From a computational perspective the update method can be thought of as a multiplicative weight update algorithm, and is shown to rapidly converge to a solution of the constraints. From a biological perspective, the model falls within the well-established framework of APC-based self-tolerance models, but introduces the key features of cross-reactivity and T cell cooperativity. The model produces the desirable features of maintaining self-reactivity within a predefined threshold, while driving the development of a diverse repertoire which can respond effectively to a broad selection of non-self antigens. The model also reproduces the heterogenous distribution of naive T cell clonotype abundances which has been described by recent high throughput sequencing studies [120], and the extensive cross-reactivity which is another recently recognized feature of the T cell repertoire [159]. We do not present a model of an immune response in this work. If the APC remains in tolerogenic state, the introduction of new non-self pMHC profiles will violate the constraints, but this will result in additional T cell killing and the system will gradually readjust to remain within the immune activation threshold. If, however, the APC are switched to an immunogenic state (for example by exposure to innate immune danger signals) then crossing the threshold will result in activation of all APC bound T cells, resulting in an effector immune response.

The mechanisms whereby the vertebrate adaptive immune system avoids harmful reaction with self antigens but retains the ability to react with a large and unknown set of potential pathogens has been extensively discussed. The current molecular understanding of the stochastic recombination events which generate adaptive immune receptors

(antibody and the TCR) require self-tolerance to be learnt rather inherited. The clonal deletion model of Burnet [27] has remained the dominant paradigm for many decades. In the context of the T cell, this paradigm posits that T cells developing in the thymus die if they react with antigens (which in the context of the thymus are assumed to be predominantly self) with an affinity above a given threshold, whose value has been estimated to correspond to a disassociation constant of approximately $6\mu M$ [115]. Indeed special molecular mechanisms exist to ensure atopic expression of a whole range of non-thymic proteins in the thymus [9], presumably to ensure robust self tolerance. The molecular mechanism of clonal deletion has also been studied intensively [61].

More recently, a number of immunologists have proposed the need for some form of extrathymic (peripheral) tolerance, since self-reactive mature T cells have been described in many cases. Such models include those in which self/non-self discrimination was assigned to the antigen presenting cell (typically a dendritic cell) rather than the T cell [95, 65]. The essence of these models was to propose that APCs exist in two different functional states. Under resting conditions (e.g. in the absence of infection) the interaction between antigen on the APC and cognate T cell induces tolerance (either deletion, or anergy). When the APC is activated (typically via the innate immune system), the same interaction leads to activation, differentiation and T cell effector function. A fundamental feature of these models is that the APC continues to present self-antigens in both states. However, since the immune system has been 'educated' to tolerise self-reactive T cells during a resting period, and the majority of antigen presenting cells at any time continue to remain in a resting state, the T cell response to self antigens presented together with non-self by the activated antigen presenting cells is small and transitory, and does not lead to significant pathology. The model presented in this paper lies squarely within the conceptual framework of these antigen presenting cell focused models of self/non-self discrimination. However, our model simplifies the system by assuming only a single type of APC. In reality, the immune system contains a heterogeneous mixture of antigen presenting cells, with a spectrum of tolerizing or activating activity [105]. The extension of our model to incorporate antigen presenting cell heterogeneity will be an important goal of future work.

The molecular mechanisms by which antigen presenting cells induce tolerance remains

an open question. Tolerogenic dendritic cells which express granzyme and perforin, and induce T cell death in an antigen specific way, have been described [166]. Dendritic cells also express several members of the Tumour Necrosis Factor (TNF) family, and its cognate receptors, the TNF receptor family. Some members of this family, for example CD40 and CD40L, are known to play a critical part in T cell activation. Impairment of this interaction leads to profound immunodeficiency [29]. Furthermore, CD40 expression on antigen presenting cells is modulated by T cells, and the antigen presenting cell integrates signals from multiple T cells, providing a molecular mechanism for T cell cooperativity [2]. Other members of the family which can be expressed by dendritic cells, in contrast, deliver negative signals. The most well-studied example is the Fas/FasL interaction, and impairment of this interaction leads to a breakdown of self-tolerance [67, 109, 162]. TNF itself can also induce cell death via TNF receptor signals, although paradoxically it can also induce cell activation [64]. The precise function of many of the more than 40 members of these families remain unknown, and their potential role in tolerance induction remains to be explored.

An interesting feature of our model is that it imposes a homeostatic limit on the total number of T cells which depends on the self-tolerance threshold. There is extensive experimental evidence linking T cell homeostasis to inter-clonal competition for the survival/proliferation cytokine IL7 [143]. An important challenge will be to integrate the phenomenon of clonal competition for a limited resource into our model. Indeed, it is possible to retain the computational infrastructure of our model but recast it emphasizing survival factors, rather than death signals. It may be the case that integration occurs in both APC and T cells, with the APC sending survival signals to bound T cells until a threshold level of binding is violated, at which point the survival signals cease. T cells would integrate the amount of survival signal received over a number of TCR-APC interactions and if this does not reach a sufficient level would die. This mechanism would increase the specificity of clonotype size adjustment, only reducing those clonotypes that repeatedly encounter APC for which the binding threshold is violated.

Of necessity, both our basic model and its implementation make a number of simplifying assumptions. The impact of some of these could be explored further by in silico

experimentation. For example, it would be relatively straightforward to implement a model in which the proliferation of the T cells is likely to be dependent on the strength of the receptor/pMHC interaction. A more complex, but important, question to explore is the extent to which the averaging of the response over all antigen presenting cells adequately captures the real scenario, where self-tolerance must be distributed anatomically over the whole body, and where each antigen presenting cell only presents a subset of all possible self antigens.

Our model does not incorporate regulatory T cells, which are clearly an important part of the mechanisms of self-tolerance, and have been the basis for several previous theoretical models of self-tolerance [31, 6]. These cells may be of particular importance for regulating those T cells with the highest affinity for self, which will still exist albeit at reduced numbers in our model, and which could be inadvertantly triggered in the context of responses to non-self with potential pathogenic consequences.

The model we propose has interesting implications for inducing organ specific tolerance in the context of allo-transplantation, which remains an unsolved problem in the context of clinical transplantation. The natural mechanisms which maintain tolerance to self are clearly insufficient in most cases to re-establish complete and lasting tolerance to an allograft in the absence of immune-suppression. This is perhaps not surprising since extra-thymic tolerance is only one component of tolerance, and in isolation may be insufficient. However, with better understanding of the molecular cell biology of tolerogenic dendritic cells, it may be possible to experimentally increase the activity or number of these cells and thus re-educate the peripheral repertoire towards tolerance.

In conclusion we propose a model of self-tolerance which incorporates T cell co-operativity (‘quorum-sensing’) into the mechanism for balancing self-tolerance with immuno-competence. Once a stable repertoire has been produced, we imagine that on immune challenge individual groups of antigen presenting cells are switched into an activated state, where they present antigens and drive the establishment of effector and memory cells. However, the repertoire will have learnt tolerance and hence the response to self will be small and not pathogenic. A useful feature of the model is that the threshold for self reaction can be set locally, and hence may vary in different tis-

sues. The balance between response and tolerance may therefore be dependent on the local micro-environment. The key prediction of our model is that perturbation of either the existing T cell repertoire, or the presented pMHC landscape will cause widespread distributed changes to the overall repertoire which will involve clones of many different specificities. The nature of these changes can be predicted by our model, and can be measured using the power of high throughput sequencing of TCR repertoires. Thus our model will stimulate further hypothesis building and falsification, and lead to a better understanding of adaptive immunity and self tolerance.

Model improvements

The implementation of the model described in this chapter has a number of limitations, including clone sizes being represented by real numbers (rather than integers) and lack of ability to model an immune response explicitly. In order to address these points, an agent based implementation of the model is being developed. The agent based model (ABM) will track each dendritic cell and each T cell individually, avoiding the problem of clone sizes becoming arbitrarily small. Additionally, it will include integration of signal on the part of the T cell to allow for positive selection to be included in a more flexible manner than the current implementation. The ABM will be used to model the situation where APCs are activated by the innate immune system, presenting a new 'foreign' peptide (along with existing self peptides) to the T cell population and causing the activation of T cells that recognise presented peptide. This will allow questions regarding the extent of activation of self-responsive T cells under an immune challenge to be explored.

Chapter 4

Heterogeneity in the antigen-specific T cell response at the level of the T cell receptor

4.1 Introduction

Much of the field of mouse immunology relies on the use of an adjuvant, in addition to the antigen of interest, to initiate a T cell response against the antigen. A commonly used adjuvant, Complete Freund's Adjuvant (CFA), consists of inactivated *M.Tuberculosis* (TB) emulsified in oil. Since T cell activation requires interactions with antigen presenting cells providing signal from peptide-MHC via their TCR as well as costimulatory signals, administering antigen alone to mice does not provoke a strong immune response. The presence of adjuvant (particularly the bacterial components) facilitates effective uptake of antigen by macrophages as well as initiating costimulatory signals in the antigen presenting cells as discussed in [69], allowing a T cell response to be studied.

Previous work from our lab [144] demonstrated that changes in the TCR repertoire could be used to discriminate between the TCR repertoires of unimmunised mice and mice immunised with CFA (with or without additional model antigen). Amino acid triplets are clustered into 'codewords' according to their Atchley factor properties, and

the occurences of these codewords in samples of CDR3βs from immunised and unimmunised mice are counted. Support Vector Machines are then used to predict whether a sample is from an unimmunised or immunised mouse, and additionally what timepoint post immunisation the sample is taken at. It was shown that whether a sample is from an immunised or unimmunised mouse is able to be predicted with 100% accuracy, and the timepoint is able to be fairly well predicted for mice sacrificed at month 2, but day 5 and day 14 repertoires are indistinguishable under this method. This classification efficiency is retained if only the largest clones from each sample are used, and remarkably is also retained if only the smallest clones are used, suggesting that the immune response is distributed across clones of different sizes.

In this chapter, we extend this work to try to identify characteristics of the TCR repertoire that change on immunisation with a model antigen, to understand more about the global effects of immunisation or the TCR-level features determining recognition of a single antigen.

4.2 Methods

4.2.1 Sample collection and sequencing

All experiments were performed by Nir Friedman's group at the Weizmann Institute, Rehovot, Israel.

C57BL/6 mice were immunised with Complete Freund's Adjuvant ('CFA'), with or without an additional antigen. Control mice were left unimmunised or were mock immunised with phosphate buffered saline ('PBS'). Additional antigens used in this study were ovalbumin ('OVA'), a protein commonly used as a model antigen, and a peptide from the heatshock protein, VLGGGCALLRCIPALDSLTPANED ('p277'). After immunisation, mice were sacrificed at either early time points (day 5 or day 7), mid time points (day 10 or day 14) or a late time point (month 2). Mice sacrificed at month 2 were given a booster of Incomplete Freund's Adjuvant at day 14.

$CD4^+$ T cells were isolated from spleens and TCRβ chains from these cells were sequenced via the protocol described in [111]. Briefly, total RNA was reversed tran-

scribed with a primer specific to the TCRβ constant region, and resulting cDNA was amplified via PCR using a set of TCRVβ primers. Illumina adaptors were ligated to the product, including indexes to identify each sample, and the sequencing was performed using a Genome Analyzer II.

For some of the mice, additional splenic $CD4^+$ T cells were taken for *in vitro* culture. 3×10^6 cells per well were cultured for a week, in the presence of irradiated autologous splenocytes and an antigen. The antigens used were heat-inactivated Tb, a peptide from ovalbumin (323-339, ISQAVHAAHAEINEAGR, referred to as 'OVA'), or the p277 peptide that was also used for *in vivo* immunisations. After culture, TCRβ chains were sequenced using the same protocol as outlined above.

Details of the immunisation, time point and *in vitro* culture of each sample used in this Chapter are outlined in Table 4.1.

Immunisation	Time point	*In vitro* culture	Number
untreated	d0		8
PBS	d5		1
PBS	m2		1
CFA	d5		3
CFA	d7		2
CFA	d14		3
CFA	m2		5
CFA	d7	TB	2
CFA	d7	OVA	2
CFA+OVA	d5		3
CFA+OVA	d7		3
CFA+OVA	d14		3
CFA+OVA	m2		6
CFA+OVA	d7	TB	2
CFA+OVA	d7	OVA	3
CFA + p277	d10		6
CFA + p277	d10	TB	5
CFA + p277	d10	p277	5

Table 4.1: Details of mouse TCR repertoire samples analysed in Chapter 4.

4.2.2 Data processing

Raw sequence data was processed first by demultiplexing using the ligated index to sort reads into the appropriate samples. Reads that did not contain any of the expected

indexes in the correct position were discarded. In order to identify the TCR present in a sequence read, our lab has previously developed software called Decombinator [145], which uses the Aho-Corasick algorithm to assign V and J genes to each read based on the presence of short 'tags' defining each potential gene. Then each read is given a five-part classifier (a 'DCR') defining the TCR, consisting of:

1. an index defining the V gene
2. an index defining the J gene
3. the number of nucleotides deleted from the 3' end of the V gene
4. the number of nucleotides deleted from the 5' end of the J gene
5. a string of nucleotides present between 3' V and 5' J.

The sequence reads from the Genome Analyzer II are 50bp long, which covers the CDR3 region of the TCRβ but does not provide coverage of the gene-specific tags used by Decombinator to assign V and J genes. Instead, to assign V and J genes to each sequence read in this dataset an adaptation of Decombinator, referred to as Short Read Decombinator ('SRD') was developed, as described in [144]. Details of the SRD algorithm are found in Appendix D. Briefly, instead of searching the sequence read for any of a set of tags uniquely identifying each gene, SRD searches the sequence read for any appropriately located substring from the 3' end of each V gene or the 5' end of each J gene. Based on the substrings that are found, the algorithm assigns a gene on the basis of length of substring matches. Once V and J genes are assigned to a sequence read, the remaining fields of the DCR can be calculated.

The genes that were used by SRD to search the sequence reads were taken from IMGT [50], and were selected from IMGT/GENE-DB (http://www.imgt.org/genedb/) with the following parameters:

- Species: Mus
- Gene type: variable/joining
- Functionality: functional
- Molecular component: TR
- Locus: TRB
- Main locus: TRB

giving 22 V genes and 11 J genes (ignoring alleles) that each sequence read could be assigned to. However, three of the V genes in this set were removed from the search due to being indistinguishable in the 3' region covered by the sequence data from other V genes (details in Appendix D).

Once each raw sequence read has been assigned a DCR, or discarded if assignment isn't possible, the data was further processed to remove those that did not describe a functional TCRβ protein. A DCR was removed if any of the following applied:

1. there was a nucleotide that the sequencing machine called ambiguously (giving an 'N' in the fastq file) in the region between the V and J genes (the fifth field of the DCR)
2. the nucleotide sequence constructed from the DCR was out-of-frame. For TCRβs, a sequence is in-frame if the 3' nucleotide of the J gene forms a codon with the two 5' nucleotides of the constant region.
3. the nucletoide sequence constructed from the DCR included a stop codon
4. the amino acid sequence translated from the DCR did not contain the conserved cysteine defining the start of the CDR3 in the correct position of the V region
5. the amino acid sequence translated from the DCR did not contain the FGXG (or related) motif in the J region defining the end of the CDR3

After processing the data files to remove non-functional TCR sequences, remaining sequences were translated and CDR3s were extracted. The analysis in this chapter mostly uses the CDR3β repertoires rather than DCR repertoires, except where gene usage is discussed.

4.2.3 Data analysis

The diversity of the CDR3β repertoires is calculated using the Simpson, Shannon, normalised Shannon and Gini indexes, as described in Section 1.3.1.1 and the Jaccard index (Section 1.3.1.3) is used to measure the similarity between pairs of repertoires.

In order to assess the distance or similarity between two CDR3β amino acid sequences, two measures are used. The Levenshtein distance counts the number of 'edits' (insertions, deletions or substitutions) that are needed to transform one of the CDR3s into

the other, while the p-spectrum kernel is a similarity measure, counting the number of substrings of length p that are shared between the CDR3s.

Other analysis techniques are described in the relevant results section. Analysis is performed in Python, using the SciPy and NumPy packages.

4.3 Characteristics of the TCR repertoire

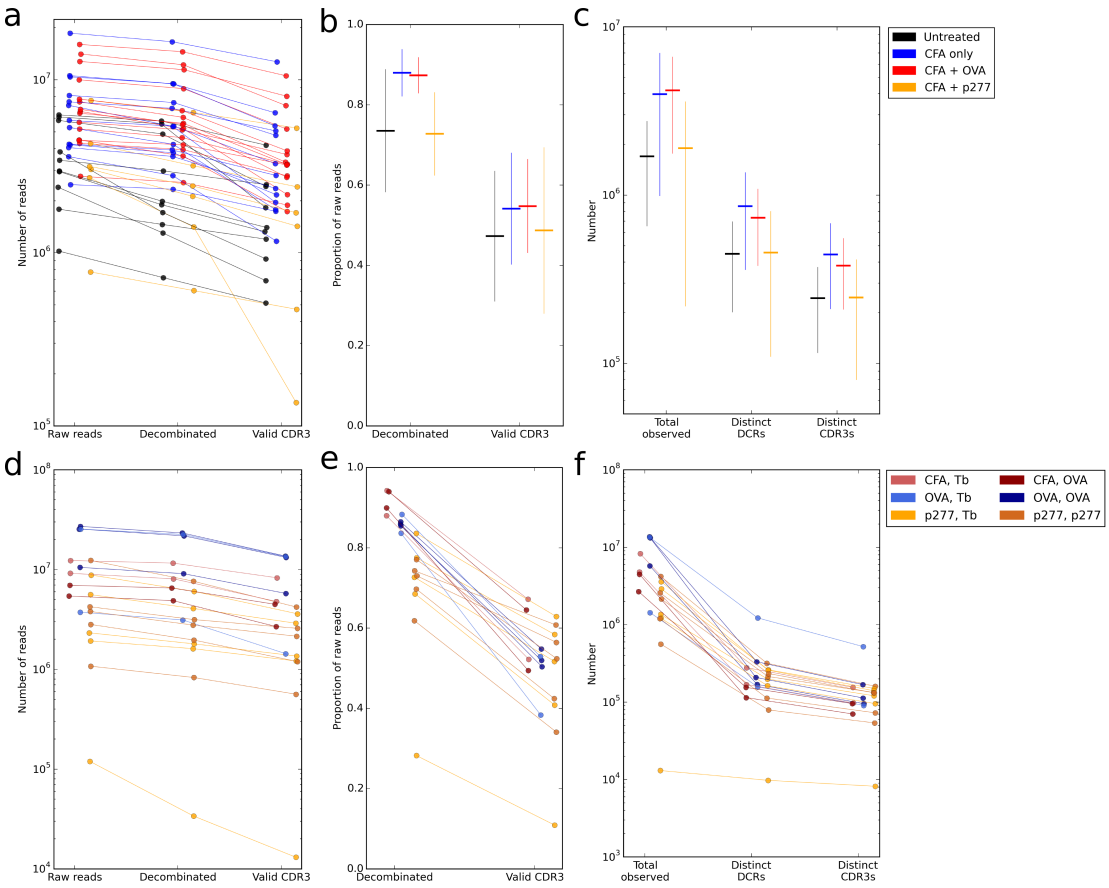

Figure 4.1: Numbers of TCR sequences obtained from each sample
For CD4$^+$ T cell receptor samples from *in vivo* immunised mice, coloured by immunisation status, the (a) number of raw reads, decombinated reads and reads with a valid CDR3 from each sample, (b) proportion of raw reads (mean $\pm$ standard deviation) that are succcessfully decombinated or have a valid CDR3 and (c) number of total observations and distinct DCRs (Decombinator classifiers) or CDR3s.
For CD4$^+$ T cell receptor samples from *in vivo* immunised mice after *in vitro* stimulation (immunisation status and *in vitro* stimulation indicated by colour), the (d) number of raw reads, decombinated reads and reads with a valid CDR3 from each sample, (e) proportion of raw reads that are successfully decombinated or have a valid CDR3 for each sample and (f) number of total observations and distinct DCRs or CDR3s per sample.

The dataset analysed in this Chapter contains 63 TCRβ repertoire samples. 44 of these

samples are from splenic $CD4^+$ T cells without any *in vitro* culture. These are referred to as *in vivo* samples, while the remaining 19 samples which are from splenic $CD4^+$ T cells after *in vitro* culture are referred to as *in vitro* samples.

The majority of the 44 *in vivo* samples contain between 10^6 and 10^7 sequence reads, and for most samples between 5×10^5 and 10^7 of these raw reads contain a valid CDR3 after processing by Decombinator (Figure 4.1a). The proportion of raw reads that are assigned a Decombinator 5-part classifier ('DCR') is lower for p277 immunised samples than unimmunised, CFA only immunised or CFA+OVA immunised samples, but the proportion of raw reads that contain a valid CDR3 is similar between immunisation groups (Figure 4.1b). The total number of TCR molecules sequenced from the *in vivo* samples (after data has been processed) is higher for CFA only and CFA+OVA immunised samples, and this difference is also seen in an increased number of distict DCRs and CDR3s observed in these samples (Figure 4.1c).

In the *in vitro* stimulated data, numbers of reads per sample are similar to the *in vivo* data, with the exception of one sample (p277 immunised and Tb stimulated) which contains only approximately 10^5 raw reads of which approximately 10^4 contain a valid CDR3 (Figure 4.1d). With the exception of this smaller sample, the proportion of raw reads that are successfully decombinated or contain a valid CDR3 (Figure 4.1e) is comparable to the *in vivo* samples, as is the number of distinct DCRs and CDR3s observed.

The proportional gene usage in the TCR repertoire samples (Figures 4.2a and b) for all samples used in this chapter shows that some genes have more variable usage between samples than others. Usage of TRBJ1-1 is highly variable, while usage of the other J genes is much more consistent. Similarly, usage of TRBV12-1 is more variable than the other V genes.

The proportional gene usage data from *in vivo* samples, grouped by immunisation status (Figure 4.2c and d) demonstrates that some genes differ between groups. However it has been shown previously in our group that these differences are not enough to distinguish between unimmunised and CFA only or CFA+OVA immunised samples. There are four J genes which show statistically significant ($p < 0.01$ by the Mann Whitney

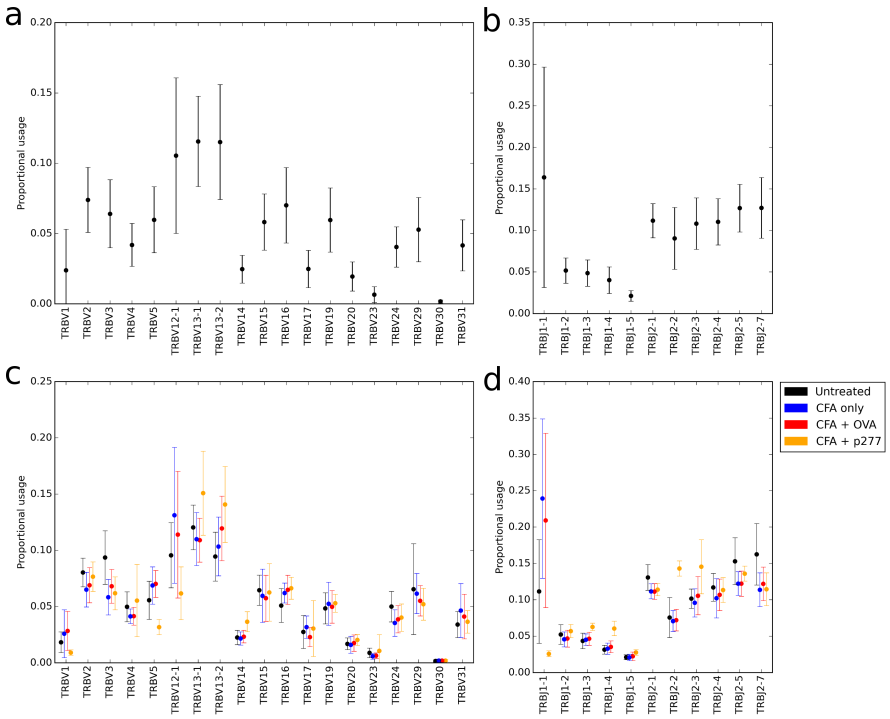

Figure 4.2: Gene usage in TCRβ samples
(a,b) Proportional V and J gene usage across all samples. (c,d) For samples from immunised mice with no *in vitro* stimulation, mean $\pm$ standard deviation of proportional V and J gene usage grouped by immunisation type (indicated by colour).

U test and the Kolmorgorov Smirnoff 2 sample test) different usage between both CFA only and unimmunised samples and between CFA+OVA and unimmunised samples (J1-1, J2-1, J2-5 and J2-6). Only one V gene (V3) demonstrates statistically different usage between CFA only and unimmunised samples and between CFA+OVA and unimmunised samples.

The gene usage profile in p277 immunised samples is markedly different from the other groups, with four J genes (J1-1, J1-3, J1-4 and J2-2) being differentially used in p277 samples when compared to any of the other groups. The V gene usage profile is less different in p277 samples, with only V5 being differentially used in p277 samples when compared to any of the other groups. However, other V genes are differentially used in p277 samples in comparison to at least one of the other samples. For example, V13-1 and V15 are both statistically significantly over-expressed in p277 samples in comparison to CFA+OVA samples.

In terms of identifying a repertoire signature of antigen-specific immune response to

ovalbumin, there are no V or J genes which have a proportional usage which is statistically significantly different between CFA only and CFA+OVA immunised groups. This indicates that the antigen specificity of the immune response cannot be identified at the level of whole repertoire gene usage.

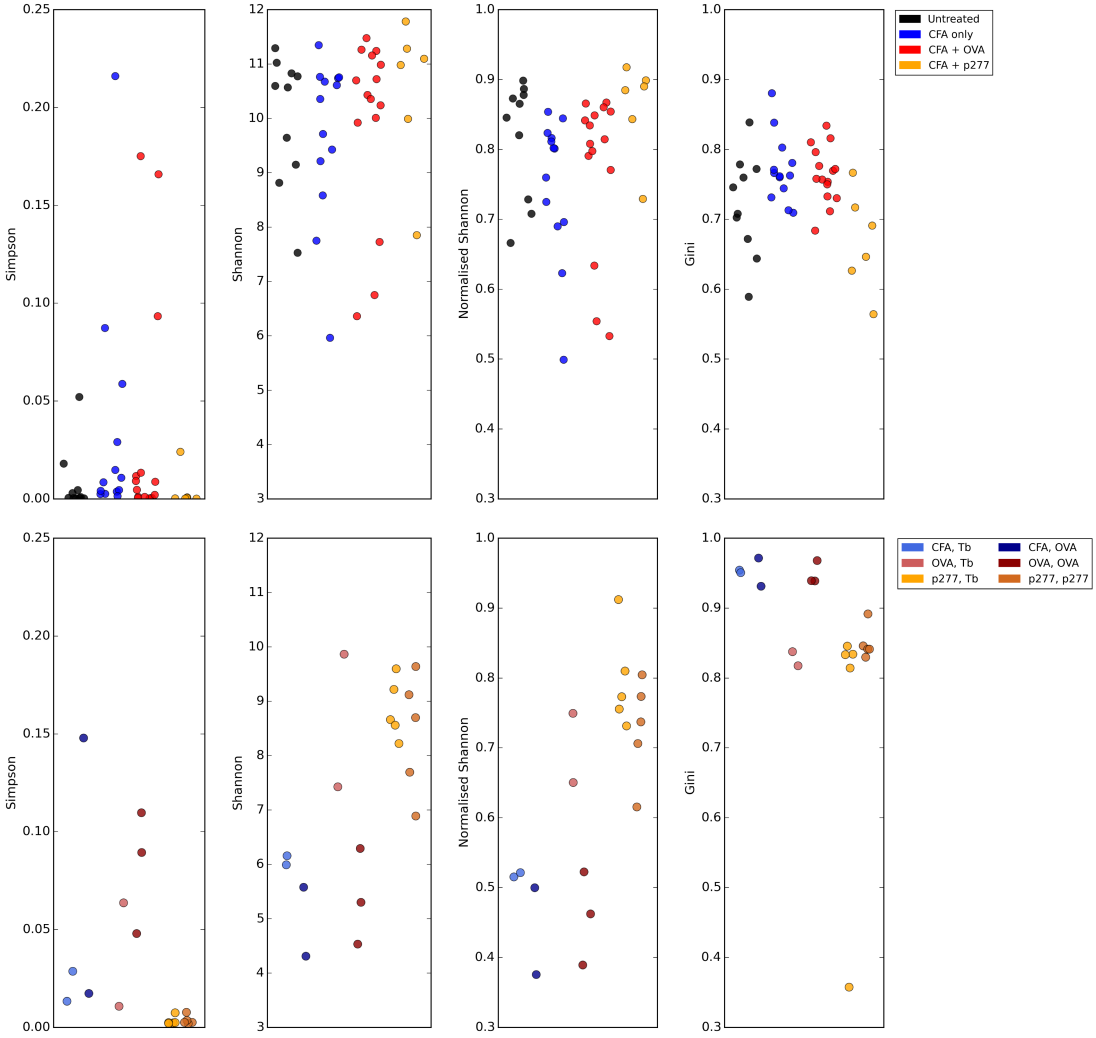

Figure 4.3: Repertoire diversity measures
The diversity of the repertoire in each sample, measured by the Simpson, Shannon, normalised Shannon and Gini indexes. Top row: diversity of *in vivo* immunised repertoires, without any *in vitro* stimulation, coloured by immunisation status. Bottom row: diversity of *in vitro* stimulated repertoires, coloured by immunisation status and *in vitro* culture peptide.

Global measures of the diversity of the repertoires (Figure 4.3) also do not show substantial differences between immunisation groups. In the *in vivo* samples, there are no statistically significant differences (by Mann Whitney U test or Kolmorgorov Smirnoff 2 sample test, with significance being $p < 0.01$) between the groups when diversity is

measured by the Simpson or Shannon indexes. By the normalised Shannon, p277 immunised samples have a higher diversity than CFA only immunised samples under the Mann Whitney U test, but not under the Kolmorgorov Smirnoff test. The Gini index, measuring the evenness of the repertoires, shows that p277 immunised samples have a lower Gini (more even distribution of clones) than other immunised samples (CFA only or CFA+OVA). None of these measures of diversity show any difference between untreated and CFA only or CFA+OVA immunised samples, and no difference between CFA only and CFA+OVA immunised samples.

For the *in vitro* samples, assessment of differences between groups is made difficult by the number of samples per group. No statistically significant differences are found between groups when samples are grouped if they are concordant in both *in vivo* immunisation status and *in vitro* stimulating peptide. When samples are grouped by either the *in vivo* immunisation status or the *in vitro* stimulating peptide then some differences are seen, although none of the groups demonstrate consistently different diversity across all measures.

The *in vitro* stimulated repertoires of p277 immunised samples demonstrate statistically significant lower diversity than CFA only or CFA+OVA immunised *in vitro* stimulated samples when diversity is measured by the Simpson index, but when diversity is measured by the Shannon or normalised Shannon this pattern is reversed and p277 immunised *in vitro* stimulated samples have higher diversity than CFA only immunised samples. p277 immunised *in vitro* stimulated samples also have a lower Gini (indicating a more even repertoire) than CFA only immunised samples.

When samples are grouped by the *in vitro* stimulating peptide, samples stimulated with p277 demonstrate lower diversity than OVA stimulated samples by the Simpson Index but higher diversity by the Shannon and normalised Shannon. When OVA stimulated and TB stimulated samples are compared using the Mann Whitney U test, OVA stimulated samples demonstrate higher diversity under the Simpson Index, but lower diversity under the Shannon and normalised Shannon. Samples stimulated *in vitro* with p277 have a lower Gini (more even repertoire) than samples stimulated with OVA.

Applying diversity metrics to these samples results in a complicated and inconsistent

picture of how immunisation and *in vitro* stimulation affects the repertoire, made difficult to interpret by the number of samples in each group. A more detailed analysis of the characteristics and composition of the repertoire in each sample is required in order to identify signatures of antigen specific response.

4.4 TCR repertoire changes on *in vivo* immunisation

The $CD4^+$ TCR repertoire samples from spleens of mice left unimmunised or immunised with CFA $\pm$ OVA are analysed to ask whether there is a set of CDR3βs that identify the immune response to OVA.

4.4.1 Known OVA-responsive CDR3βs do not distinguish repertoires from OVA-immunised mice

We measure the frequency of known OVA-responsive CDR3βs in each of our samples. The OTII T cell is a $CD4^+$ T cell expressing a TCR known to be responsive to the OVA peptide in C57BL/6 mice (e.g. [123]), and the frequency of this CDR3β is increased in immunised mice (Figure 4.4a). However, the frequency is not further increased in mice which have been immunised with CFA+OVA rather than CFA only or CFA + p277. This is not an artefact of sample size, since the same pattern is observed in absolute abundances. Additionally, the distribution of frequencies across all observed CDR3βs and the number of CDR3βs observed is not substantially altered by immunisation, suggesting that the increased frequency observed on immunisation is due to clonal expansion of cells expressing the β chain of OTII, but not in an OVA-specific manner.

The OTI T cell is a $CD8^+$ T cell with a TCR recognising a peptide of ovalbumin. The CDR3β of OTI is observed at low frequency in some of our samples (Figure 4.4b), suggesting that either OTI is also present on some $CD4^+$ T cells or that the experimental sorting of $CD4^+$ from $CD8^+$ T cells was not completely pure. The frequency of OTI CDR3β is not increased in immunised mice. The OTII CDR3α is not observed in any sample in this data, nor is the DO11.10 CDR3β, a TCR responsive to ovalbumin in the context of I-Ad, an MHC molecule not expressed in the mice in our experiment.

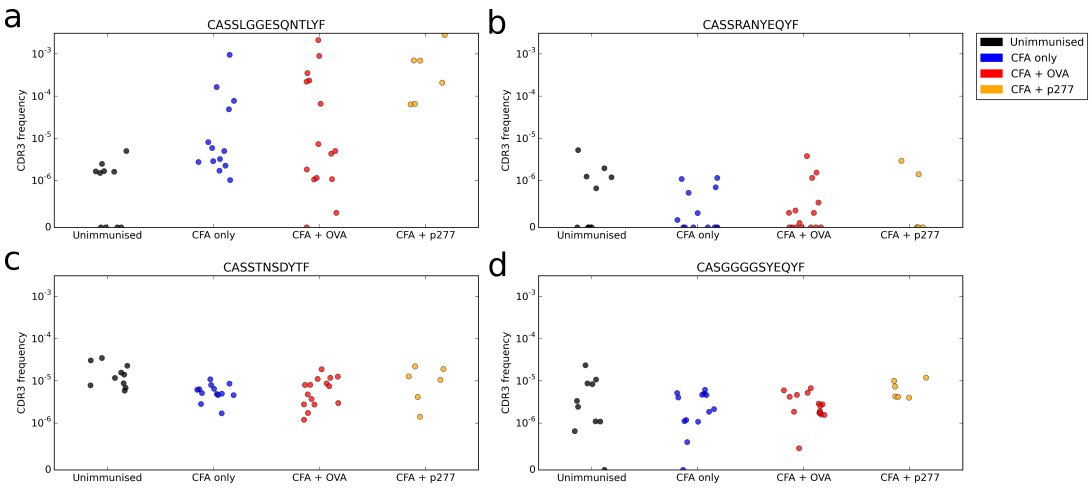

Figure 4.4: Frequency of known OVA-responsive CDR3s: *in vivo* samples
The frequency in TCR repertoire samples of (a) OTIIβ, (b) OTIβ, (c,d) CDR3s identified in [46] as OVA-responsive.

In a study investigating the T cell response in mice to skin sensitisation assays [46] TCRs were sequenced from skin and organs before and after applying ovalbumin (with cholera toxin as adjuvant) to the ear. A number of CDR3s are identified which are undetectable before sensitisation and highly abundant in skin and lymph nodes once acute inflammation is resolved. Of these OVA-responsive CDR3s identified, two are present in our samples (Figure 4.4c,d) but again are not observed to be increased in mice that have been immunised.

4.4.2 OVA-responsive TCRs are not public between mice

Since known OVA-responsive CDR3βs do not differentitate between samples from mice immunised with or without ovalbumin, it seems that the antigen-specific TCR response might be distributed amongst many different CDR3βs. We hypothesise that these CDR3βs might be shared between mice, and therefore samples from mice that have been subject to immunisation with the same antigens might have more CDR3βs in common than samples from mice who have not been subject to the same immunisation.

There is no increase in number of CDR3βs that are highly public in groups of immunised mice in comparison to unimmunised mice (Figure 4.5a), suggesting that the T cell response to immunisation does not involve the same CDR3βs in all the mice. Next, the convergence of immune response between pairs of mice is considered using the Jaccard

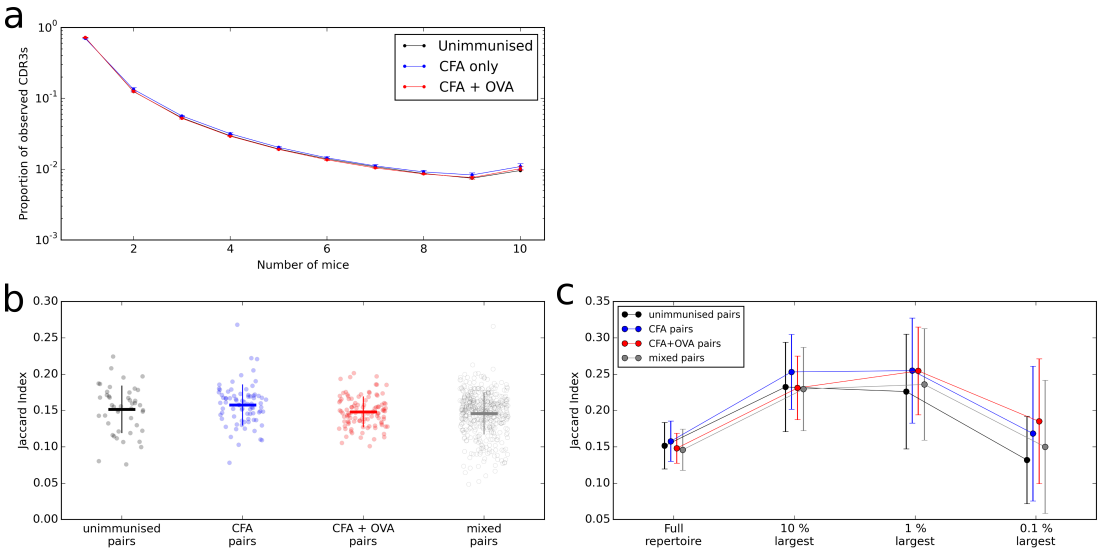

Figure 4.5: Immune response at the CDR3β level is not shared between mice
(a) The proportion of CDR3βs that are observed in exactly the number of mice on the x-axis, with mice grouped by immunisation status. Equal sized subsamples of mice are repeatedly selected from each group, with mean $\pm$ standard deviation shown.
(b) The Jaccard index between the spleen sample CDR3β repertoires from every pair of mice, grouped by immunisation status. Lines indicate mean $\pm$ standard deviation for each group.
(c) The mean $\pm$ standard deviation Jaccard Index for each group, as in (b), when the repertoire considered is restricted to the largest CDR3βs observed in each sample.

Index. The Jaccard index is a measure of similarity of two sets, and here measures the number of distinct CDR3βs that are common to both repertoires as a proportion of the number of distinct CDR3βs that appear in either repertoire.

The distribution of pairwise Jaccards between repertoires from unimmunised mice or immunised mice, either under the same or different immunisation conditions (Figure 4.5b) does not suggest any convergence of repertoire on immunisation can be detected at the full repertoire level. Next, a restricted repertoire for each sample, including only those CDR3βs with the largest abundances in the sample, is considered. Restricting the repertoire in this manner increases the absolute Jaccard index between samples when considering the 10% or 1% most abundant CDR3βs (Figure 4.5c), showing that moderately frequent CDR3βs are more public than rare CDR3βs and very highly frequent CDR3βs. However, there is no statistically significant difference between the distribution of Jaccards from unimmunised and immunised pairs of mice (also seen in the same data set in [91]), suggesting that this increased absolute value of the Jaccard

is not due to convergence of the repertoire on immunisation. Taking only the 0.1% most abundant CDR3βs does give a statistically significant increase in the distribution of pairwise Jaccards on immunisation ($p < 10^{-3}$ for unimmunised vs CFA and unimmunised vs CFA+OVA under both the Mann Whitney U test and the Kolmorgorov-Smirnoff 2 sample test). This suggests that in the very largest CDR3βs there may be convergence of repertoire due to immunisation, but no convergence in response to the OVA element of the immunisation.

4.4.3 Expansion with respect to unimmunised samples identifies privately responding CDR3βs

The most abundant CDR3βs are not necessarily those that have undergone the most clonal expansion on immunisation and therefore they might not demonstrate a convergent immune response. In order to identify the most highly proliferated CDR3βs, the frequency observed in a sample needs to be compared to a 'background' frequency. Here the background frequency is defined as the frequency of the CDR3β when all observations from unimmunised mice are pooled. If a CDR3β is found in none of the unimmunised samples it is allocated a background frequency equal to the minimum background frequency observed.

Expansion coefficients for each CDR3β in each immunised mouse sample are calculated by dividing the CDR3β frequency in the immunised sample by the CDR3β background frequency, obtained as described above. The proportion of observed CDR3βs in each sample that have expansion coefficient greater than a given threshold is calculated, with means for each immunisation group shown in Figure 4.6a. There is no statistically significant difference between the proportion of CDR3βs expanded in CFA only and CFA+OVA immunised group at any of the expansion thresholds (by the KS 2 sample test, all p-values > 0.01).

For the following analysis we define 'highly expanded CDR3βs' in a sample from a mouse as those that have an expansion coefficient greater than 2^9. The proportion of CDR3βs from each sample that are defined as highly expanded does not differ between CFA only and CFA+OVA immunised groups (Figure 4.6b). There is also no statisti-

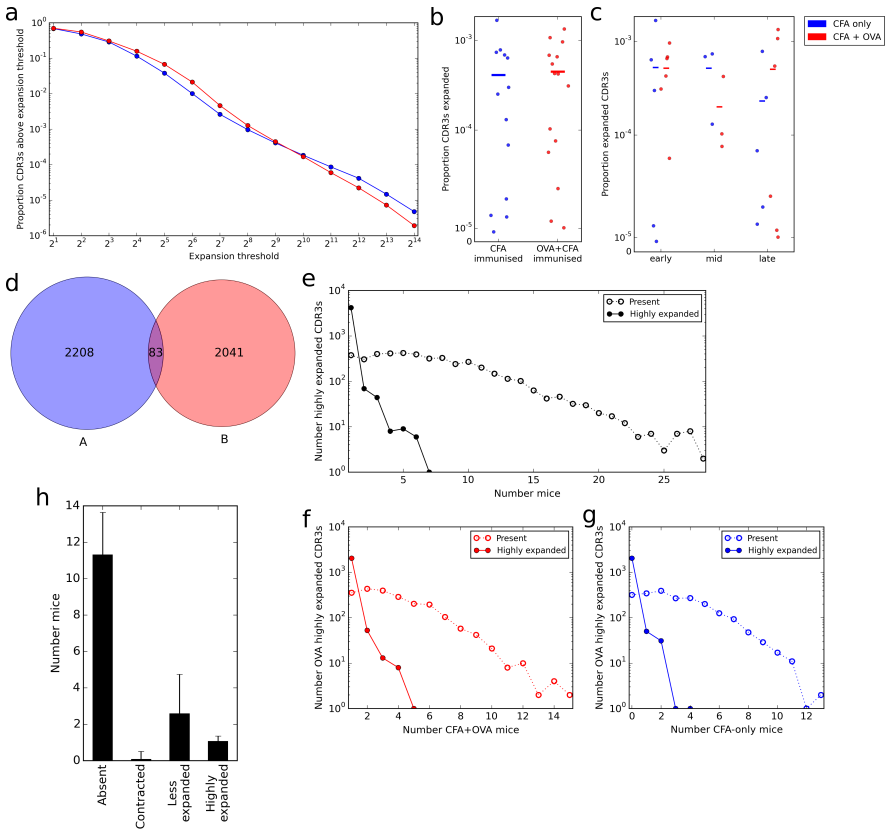

Figure 4.6: Expanded CDR3βs identified with reference to frequency in unimmunised mice

For each CDR3β in a sample from an immunised mouse, an expansion coefficient relative to the frequency in pooled samples from unimmunised mice is calculated. (a) The proportion of observed CDR3βs in immunised samples with expansion coefficient greater than expansion thresholds on the x-axis. Data show is mean across all mice from an immunisation condition and colour indicates immunisation condition. Using a definition of expanded CDR3βs as those with an expansion coefficient in the sample $> 2^9$, the proportion of CDR3βs in each sample defined as expanded, with samples grouped by (b) immunisation condition or (c) immunisation condition and time of sample extraction. Early: d5 or d7 post immunisation, mid: d10 or d14 post immunisation, late: m2 post immunisation. Colour indicates immunisation condition. (d) Venn diagram demonstrating overlap between sets of highly expanded CDR3βs. Set A: The CDR3βs highly expanded in at least one CFA-only immunised mouse. Set B: The CDR3βs highly expanded in at least one CFA+OVA immunised mouse. (e) For CDR3βs defined as highly expanded in any mouse (total 4,332), the number of mice (from either immunisation group) that they are present (dotted line, open circles) or highly expanded (solid line, solid circles) in. For CDR3βs defined as highly expanded in at least one CFA+OVA immunised mouse (total 2,124), the number of (f) CFA+OVA immunised mice and (g) CFA only immunised mice that they are present (dotted line, open circles) or highly expanded (solid line, solid circles) in. (h) The mean number of CFA+OVA immunised mice (total 15) that a CDR3β defined as highly expanded in at least one CFA+OVA immunised mouse (total 2,124) falls into each category within. CDR3βs are categorised as being (i) absent from a sample if not observed at all, (ii) contracted if they are observed with an expansion coefficient < 1, (iii) less expanded if expansion coefficient is > 1 but $< 2^9$ and (iv) highly expanded if expansion coefficient is $> 2^9$. Error bars represent standard deviation.

cally significant difference between the CFA only and CFA+OVA immunised groups when the samples are separated according to time post-immunisation before sample collection (Figure 4.6c, by the KS 2 sample test all p-values > 0.1 for CFA-only vs CFA+OVA within a timepoint). While the 'magnitude' of the response to immunisation does not appear to depend on the presence of OVA antigen, the clonal composition of the response might alter, allowing identification of OVA-responsive CDR3βs.

The CDR3βs that are highly expanded in any CFA-only immunised mouse (2,291 total) and those that are highly expanded in any CFA+OVA immunised mouse (2,124 total) show little overlap (Figure 4.6d). This suggests an individual-specific response to CFA is being observed, since if the CDR3βs expanding in response to CFA were shared between mice the set of CDR3βs highly expanded in CFA-only mice might be expected to be a subset of those highly expanded in CFA+OVA mice. This is further demonstrated by counting the number of mice (of either immunisation group) that each highly expanded CDR3β (total 4,332) is present or highly expanded in (Figure 4.6e). Although many of these CDR3βs are present in a number of mice, the majority (4,195) are only highly expanded in a single mouse and the most publicly highly expanded CDR3β is highly expanded in just seven of 28 mice.

The CDR3βs that are highly expanded in at least one CFA+OVA immunised mouse we call 'OVA highly expanded'. The OVA highly expanded CDR3βs are also not very public among CFA+OVA immunised mice (Figure 4.6f), with 2,049 of the 2,124 being highly expanded in only one mouse and the most public OVA highly expanded CDR3β being highly expanded in just five of the 15 mice, despite some of them being present in many of the mice. Similarly, the OVA highly expanded CDR3s are not highly expanded in multiple CFA-only immunised mice (Figure 4.6g).

The clonal expansions in the T cell population in response to immunisation with OVA appear to be mostly private to each mouse. This privacy of response could be due to T cells of a responsive CDR3β not being present in multiple mice before immunisation, so not being available for expansion. Alternatively, it could be that cells bearing the CDR3β are present in multiple mice but are not selected for expansion in all mice, possibly due to differential competitive effects from the rest of the T cell population. For

each of the OVA highly expanded CDR3βs, the expansion coefficient of the CDR3β in each of the CFA+OVA immunised mice was collected and the 'behaviour' of the CDR3β in each of the mice was allocated to one of four categories. If the CDR3β isn't observed in the sample from a mouse it is classed as 'absent', while if it is present but with an expansion coefficient < 1 it is categorised as 'contracted'. If the CDR3β has an expansion coefficient > 1 it is either 'less expanded' if the expansion coefficient isn't greater than the 2^9 threshold, or 'highly expanded' if it is. For each OVA highly expanded CDR3β the number of CFA+OVA immunised mice (total 15) that it demonstrates each of these behaviours in is counted. The mean of these numbers across all OVA highly expanded CDR3βs (Figure 4.6h) shows that in very few mice is an OVA highly expanded CDR3β contracted and for the majority of mice it is completely absent from the sample. This suggests that the privacy of the clonal response to OVA is due to the expanding CDR3βs not being available in other mice.

To identify a clonal signature of response to OVA, we attempt to separate the CDR3βs that are proliferating in response to the CFA from those proliferating in response to the OVA. Initially, the OVA highly expanded CDR3βs (those with expansion coefficient $> 2^9$ in at least one CFA+OVA immunised mouse) are filtered by their maximum expansion coefficient in any CFA only immunised mouse. A lower expansion coefficient threshold is defined, and if a CDR3β is OVA highly expanded and does not exceed this threshold in any CFA only immunised mice we refer to it as 'OVA specifically highly expanded'.

The number of OVA specifically highly expanded CDR3βs from each CFA+OVA immunised mouse are counted (Figure 4.7a). This number remains relatively stable across different values of the lower expansion coefficient threshold, and when this threshold is set to 2^4 all CFA+OVA immunised mice contain at least one OVA specifically highly expanded CDR3β. However, the OVA specifically highly expanded CDR3βs are very private, with only one meeting the criteria in more than one CFA+OVA immunised mouse, and therefore may not be suitable to define a CDR3-level signature of T cell population response to OVA immunisation. We collect the total set of these OVA specificially highly expanded CDR3βs and refer to them as a set of OVA-associated CDR3βs.

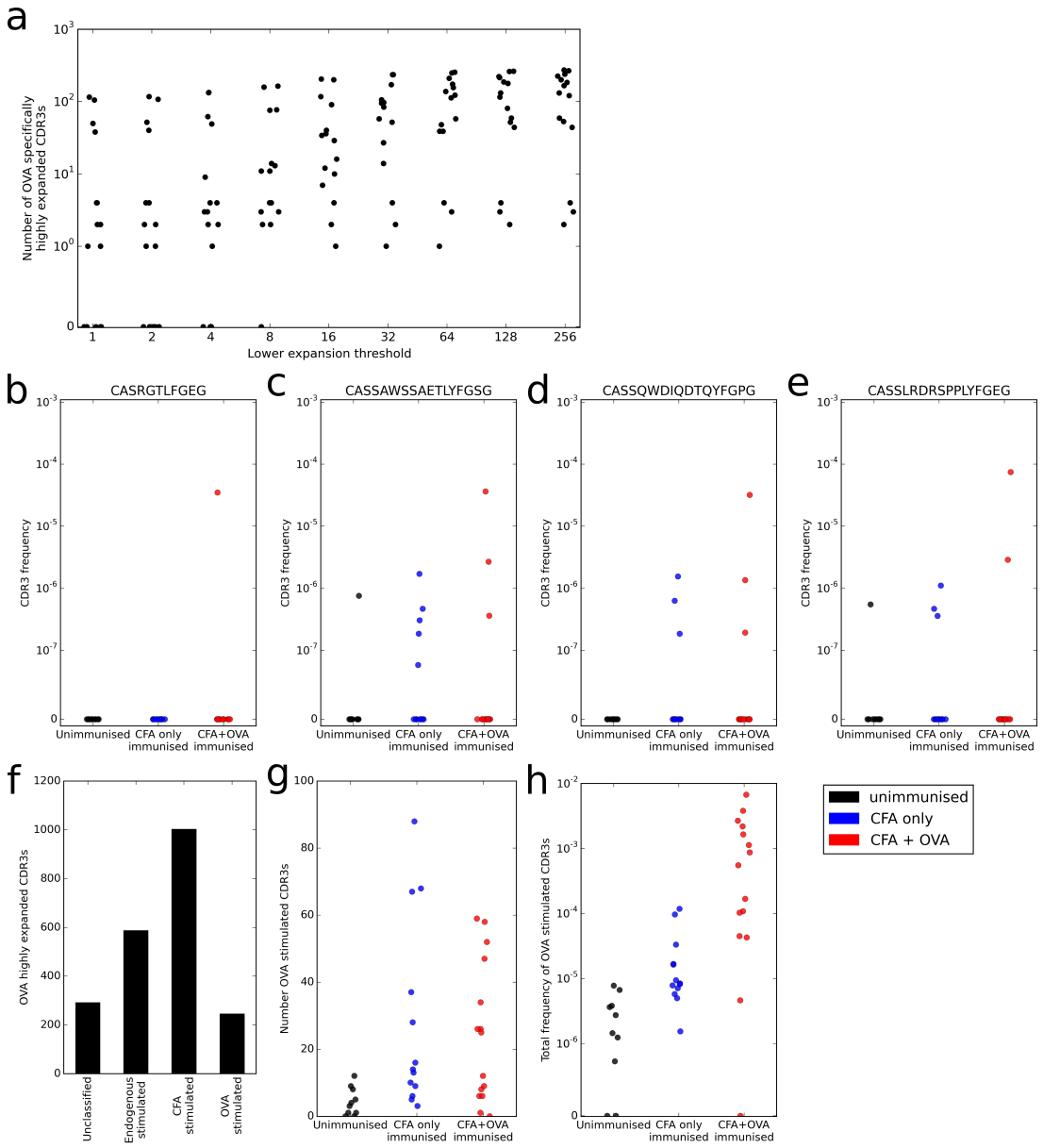

Figure 4.7: Identification of OVA-associated CDR3βs in *in vivo* data
(a) For each CFA+OVA immunised mouse, the number of CDR3βs defined as OVA specifically highly expanded at different lower expansion coefficient thresholds. (b)-(d) Representative examples of OVA highly expanded CDR3βs that are categorised as (b) unclassified, (c) endogenous peptide stimulated, (d) CFA stimulated or (e) OVA stimulated. (f) The number of OVA highly expanded CDR3βs that are classified into each of these four categories. (g) The number of OVA-associated CDR3βs present in each mouse sample. (h) The total frequency of OVA-associated CDR3βs in each mouse sample.

Next, we consider the reasons why the OVA highly expanded CDR3βs might have proliferated, and use the frequency of the CDR3β in all samples to determine which reason is most likely as follows. A CDR3β might clonally expand due to recognition of an endogenous peptide, in which case we would expect the CDR3β to be present in an unimmunised mouse at a comparable frequency to the lowest non-zero frequency observed in a CFA+OVA immunised mouse. Alternatively, it might have clonally expanded in response to a peptide from the adjuvant. In this scenario we would expect to observe the CDR3β at comparable frequency in at least one CFA only immunised mouse sample. We classify each of the OVA highly expanded CDR3βs into one of the following four categories:

Unclassified: An OVA highly expanded CDR3β which is not present in any sample from an unimmunised or CFA only immunised mouse

Endogenous stimulated: An OVA highly expanded CDR3β with maximum frequency across unimmunised mice $> 0.5\times$ minimum non-zero frequency across CFA+OVA immunised mice.

CFA stimulated: An OVA highly expanded CDR3β which is not unclassified or classified as endogenous peptide stimulated and which has maximum frequency in CFA immunised mice $> 0.5\times$ minimum non-zero frequency across CFA+OVA immunised mice.

OVA stimulated: An OVA highly expanded CDR3β which is observed in at least one sample from unimmunised or CFA only immunised mice but which is not classified as endogenous stimulated or CFA stimulated.

Representative examples of OVA highly expanded CDR3βs falling into each of these categories are shown in Figures 4.7b-e. Of the 2,124 OVA highly expanded CDR3βs, 291 cannot be classified, 586 appear to be expanded due to endogenous peptide, 1,002 are classified as CFA stimulated and 245 are classified as OVA stimulated (Figure 4.7f). These form a second set of OVA-associated CDR3βs. Despite these OVA-associated CDR3βs being highly expanded (relative to pooled unimmunised mice) in a CFA+OVA immunised sample and being present at a higher frequency in CFA+OVA immmunised

mice than any unimmunised or CFA immunised mouse, they do not seem to define a cross-animal signature of immune response. In particular, in one CFA+OVA immunised mouse none of these OVA-associated CDR3βs appear and the number of them that are present (at any frequency) in each of the CFA+OVA immunised mice is not higher than the number that are present in the CFA only immunised samples (Figure 4.7g). The total frequency of this set of OVA-associated CDR3βs in CFA+OVA immunised samples is higher than in CFA only immunised samples (Figure 4.7h), but the distributions do overlap.

4.5 *In vivo* TCR repertoires: discussion

We have seen that CDR3βs previously described as being part of an OVA-specific TCR do not have increased frequency in OVA immunised samples in comparison to CFA immunised samples in these TCR repertoires. However, they do have increased frequency in immunised samples in comparison to unimmunised samples. We propose that this observation can be explained by one of the following: (i) these β chains are paired with multiple αs giving different specificities for either CFA or OVA in different mice, (ii) these β chains create cross-reactive TCRs, responding to both the OVA peptide and a presented antigen from the adjuvant, or (iii) the cells bearing these β chains are activated via a 'by-stander' effect of some description, perhaps because of increased presentation of a particular self-peptide due to the immune response or recruitment by activated clones. It would be interesting to perform the same experiment in the context of a different adjuvant, to determine if the CFA component of the immunisations is driving the proliferation of these CDR3βs.

We also see in these data that concordant immunisation status does not increase the amount of CDR3β sharing between pairs of repertoires when the full repertoire is considered, and only slightly increases the level of sharing between repertoires when just the very largest clones are considered. This suggests that immunisation does not drive a convergent immune response at the CDR3β level even in genetically identical mice, so it will not be possible to define the TCR response to antigen challenge in terms of frequency of single CDR3βs.

We can use these TCR repertoires to identify CDR3βs that have high frequency in an immunised sample in comparison to pooled unimmunised samples, which could result from clonal expansion on immunisation. We are able to define a set of OVA-associated CDR3βs that each appear to have been highly expanded in a CFA+OVA immunised animal, but not highly expanded in any CFA-only immunised animals. These OVA-associated CDR3βs were remarkably private, with most only being highly expanded (relative to unimmunised samples) in one or two CFA+OVA immunised animals, suggesting that each animal employs a distinct set of CDR3βs in their ovalbumin response. Further, these OVA-associated CDR3βs are completely absent from the majority of other mice rather than being present but not abundant, suggesting that the privacy of the ovalbumin response is due to lack of availability of the same clone in multiple mice rather than differential selection.

The privacy of expanded clones was reinforced in analysis which defines OVA-associated CDR3βs by comparison of their minimum non-zero frequency in CFA+OVA immunised mice with their maximum frequency in other mice. This analysis obtains a set of 245 OVA-associated CDR3βs, which are again found to be predominantly privately OVA-associated and the most 'public' of these is present in only four of the CFA+OVA immunised mice. Additionally, this analysis inferred that the majority of CDR3βs that are highly expanded in a CFA+OVA immunised sample compared to unimmunised mice can be classified as stimulated by a CFA peptide or by an endogenous peptide. It may be that in this experimental set up the T cell response to the OVA peptide is dwarfed by the possibly broader and stonger response to the adjuvant. The *in vitro* stimulation data from these experiments, where CD4$^+$ T cells are cultured in the presence of TB or OVA peptide may provide more insight into the repertoire changes in response to OVA and is discussed next.

4.6 TCR repertoire changes on *in vitro* stimulation

For a number of the immunised mice, splenic CD4$^+$ T cells were cultured *in vitro* with heat-killed TB, OVA peptide or p277 for 7 days and the TCR repertoire was then sequenced. These samples are used here to identify CDR3βs which are potentially

responsive to ovalbumin.

4.6.1 Known OVA-responsive CDR3s are not increased in frequency in OVA stimulated *in vitro* cultures

The frequency of OVA-responsive CDR3s described in Section 4.4.1 is measured in the *in vitro* cultured samples and the fresh spleen samples of the same mice (Figure 4.8). OTIIβ is found at relatively high frequency in all samples (Figure 4.8a), but is not increased in OVA stimulated samples. OTIβ (Figure 4.8b) is not observed in many *in vitro* stimulated samples and is at low frequency when observed, similar to the pattern in the *in vivo* samples (Figure 4.4b). The OVA-responsive CDR3s identified in [46] (Figure 4.8c,d) are also not increased in frequency on *in vitro* stimulation with OVA.

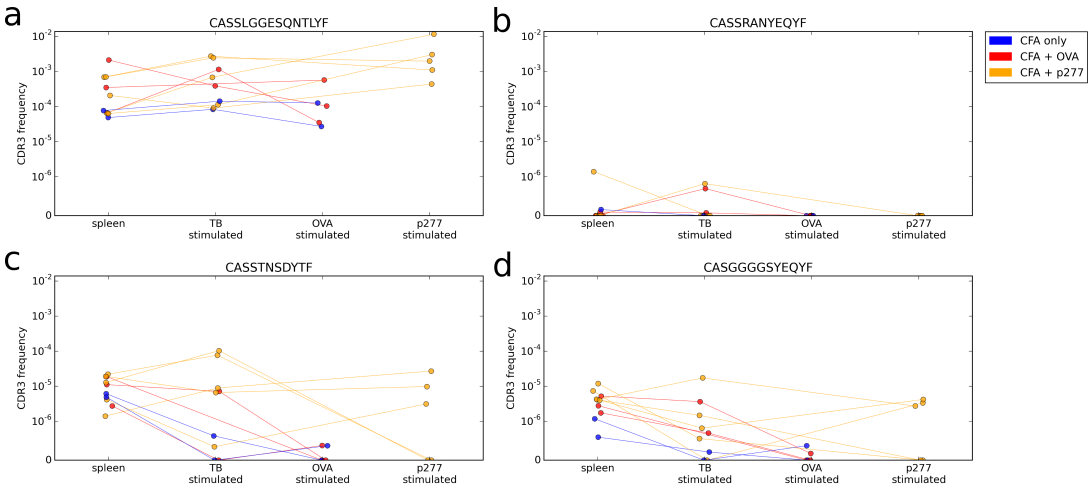

Figure 4.8: Frequency of known OVA-responsive CDR3s: *in vitro* stimulated samples The frequency in TCR repertoire samples of (a) OTIIβ, (b) OTIβ, (c,d) CDR3s identified in [46] as OVA-responsive. Colour indicates the *in vivo* immunisation treatment of the mouse. Values in the spleen column are the frequency in the sample sequenced directly from the spleen, other columns are the frequency in the sample sequenced after *in vitro* stimulation for 7 days with the indicated antigen. Lines join samples derived from the same animal.

4.6.2 *In vitro* stimulation of pools of spleen cells with different antigens

Within the dataset there are four mice for which there are sequenced TCR samples taken directly from the spleen, after *in vitro* culture with TB and after *in vitro* culture with OVA. These samples are used to search for OVA-responsive CDR3βs. In this analysis

these mice are referred to as A, B, C and D. Mice A and B were *in vivo* immunised with CFA only, and sacrificed at d7 post immunisation. Mice C and D were *in vivo* immunised with CFA+OVA and were also sacrificed at d7 post immunisation. There are a total of 12 samples sequenced from these four mice.

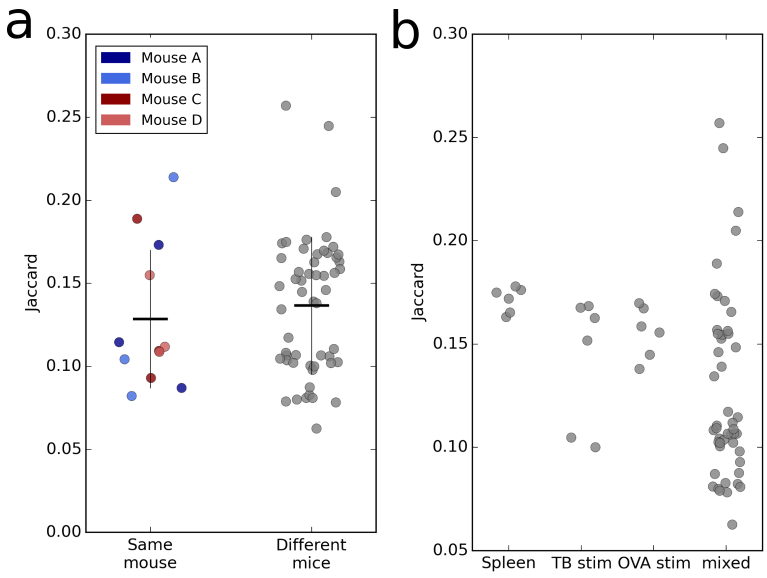

Figure 4.9: Similarity between repertoires
The similarity, as measured by the Jaccard index, between pairs of CDR3β repertoires from four mice with spleen, OVA stimulated and TB stimulated samples when pairs are grouped by (a) whether they are from the same or different animals and (b) whether they are from the same or different sample types.

For any pair of these 12 samples, there is no increased similarity (as measured by the Jaccard Index) between the CDR3β repertoires when the samples are from the same mouse (Figure 4.9a). Additionally there is no increased similarity between pairs of samples of the same sample type (spleen, TB stimulated or OVA stimulated) when compared to samples of different types (Figure 4.9b).

4.6.3 Expansion relative to *ex vivo* spleen sample identifies privately OVA associated CDR3βs

For each of these 12 samples, we consider the distribution of CDR3β clone sizes (Figure 4.10a-c). In the samples sequenced directly from *ex vivo* spleen (Figure 4.10a) the CDR3β abundances are comparably distributed, regardless of *in vivo* immunisation condition. In the TB stimulated samples (Figure 4.10b), the distributions are less

similar. Three of the four samples contain fewer distinct CDR3βs than the spleen samples, and the most abundant CDR3βs from mouse D have smaller observed clone size than the most abundant CDR3βs from the other mice. In the OVA stimulated samples (Figure 4.10c) again fewer distinct CDR3βs are observed when compared to spleen samples. Additionally, the samples from CFA+OVA immunised mice (red lines) appear to have more abundant CDR3β clones than samples from CFA only immunised mice after *in vitro* OVA stimulation.

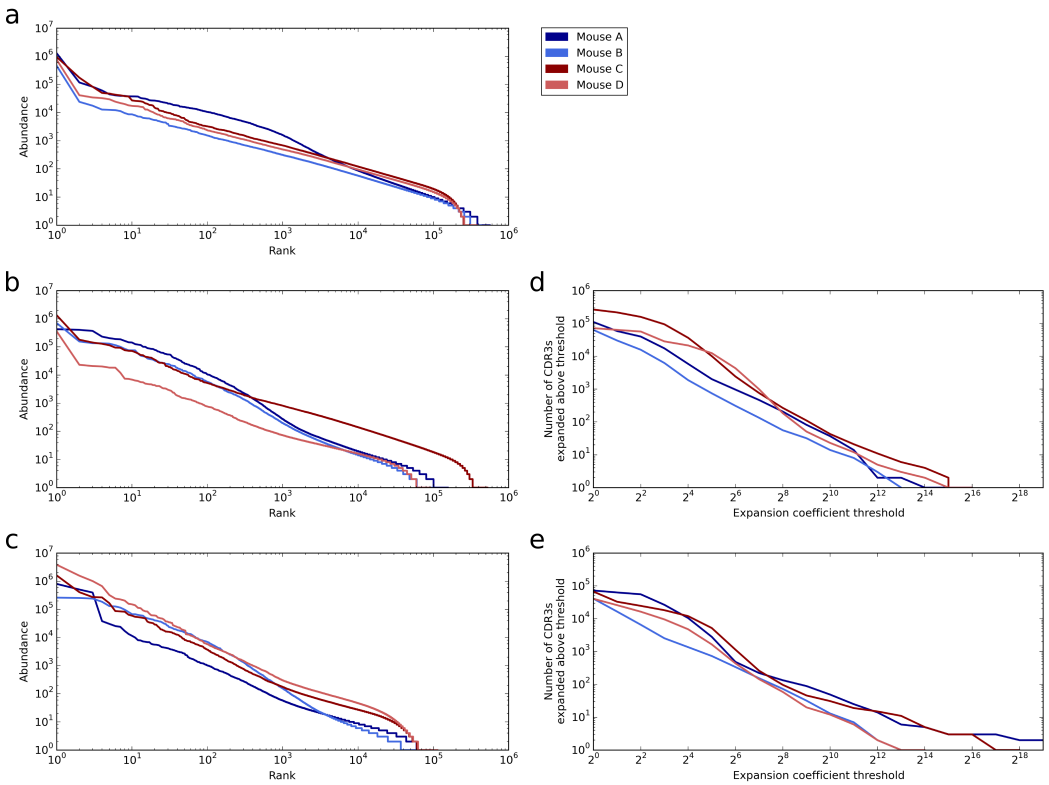

Figure 4.10: CDR3β clone sizes and expansion coefficients of *in vitro* samples
Mice A and B immunised with CFA only (blue), mice C and D immunised with OVA + CFA (red). The abundance of ranked CDR3βs from each mouse in (a) *ex vivo* spleen samples, (b) TB stimulated samples and (c) OVA stimulated samples. Expansion coefficients for each CDR3β in the stimulated samples are calculated by dividing their frequency by their frequency in the *ex vivo* spleen sample from the same animal. The number of CDR3βs from each mouse that have an expansion coefficient greater than the indicated threshold in (d) TB stimulated samples and (e) OVA stimulated samples.

For each CDR3β observed in an *in vitro* stimulated sample, an expansion coefficient can be calculated by dividing its frequency by the frequency of the same CDR3β in the *ex vivo* spleen sample from the same animal. If the CDR3β is not observed in the spleen sample it is assigned the minimum observed frequency to use as a denominator

to calculate the expansion coefficient. The expansion coefficient estimates the proliferation of the CDR3β due to the *in vitro* stimulation condition, and does not include the proliferation of the CDR3β due to the immunisation that the animal was given *in vivo*.

An expansion coefficient is calculated in this way for every CDR3β observed in each *in vitro* stimulated sample. The number of CDR3βs that have expansion coefficient greater than a given threshold in the TB stimulated samples (Figure 4.10d) and in the OVA stimulated samples (Figure 4.10e) are broadly similar for each mouse. In a similar manner to the analysis of the *in vivo* data, we define OVA-associated CDR3βs for each mouse as those with a high expansion coefficient in the OVA stimulated sample and a low expansion coefficient in the TB stimulated sample.

The relationship between the expansion coefficient of a CDR3β in the TB and OVA stimulated samples from a mouse are shown in Figure 4.11. In each of the mice there is statistically significant correlation in the expansion coefficients (Pearson correlation, $p < 0.002$ for each mouse). However, if only the CDR3βs that are expanded at least 8 fold in one of the *in vitro* stimulated samples relative to the spleen sample are included in the analysis the correlation becomes non-significant ($p > 0.05$) for mice C and D while remaining highly significant for mice A and B. This suggests that after immunisation with CFA only (mice A and B), CDR3βs proliferate equivalently on *in vitro* exposure to TB or OVA. However, after immunisation with CFA+OVA (mice C and D), subsequent *in vitro* exposure to OVA stimulates a different set of CDR3βs than are proliferated on exposure to TB.

We define OVA-associated CDR3βs in each of these mice as those that fall into the top left quadrant of the plots in Figure 4.11, that is, CDR3βs which have an expansion coefficient $< 2^5$ on exposure to TB but $> 2^9$ on exposure to OVA. These criteria define 75 OVA-associated CDR3βs in mouse A, 2 in mouse B, 38 in mouse C and 9 in mouse D. This gives a total set of 124 OVA-associated CDR3βs derived from the *in vitro* sequence data, each of which is completely privately OVA-associated, i.e. it is only defined as OVA-associated in one of the four mice.

While these 124 CDR3βs are privately OVA-associated, they are not completely private. Figure 4.12a shows the abundance of each of the 124 OVA-associated CDR3βs,

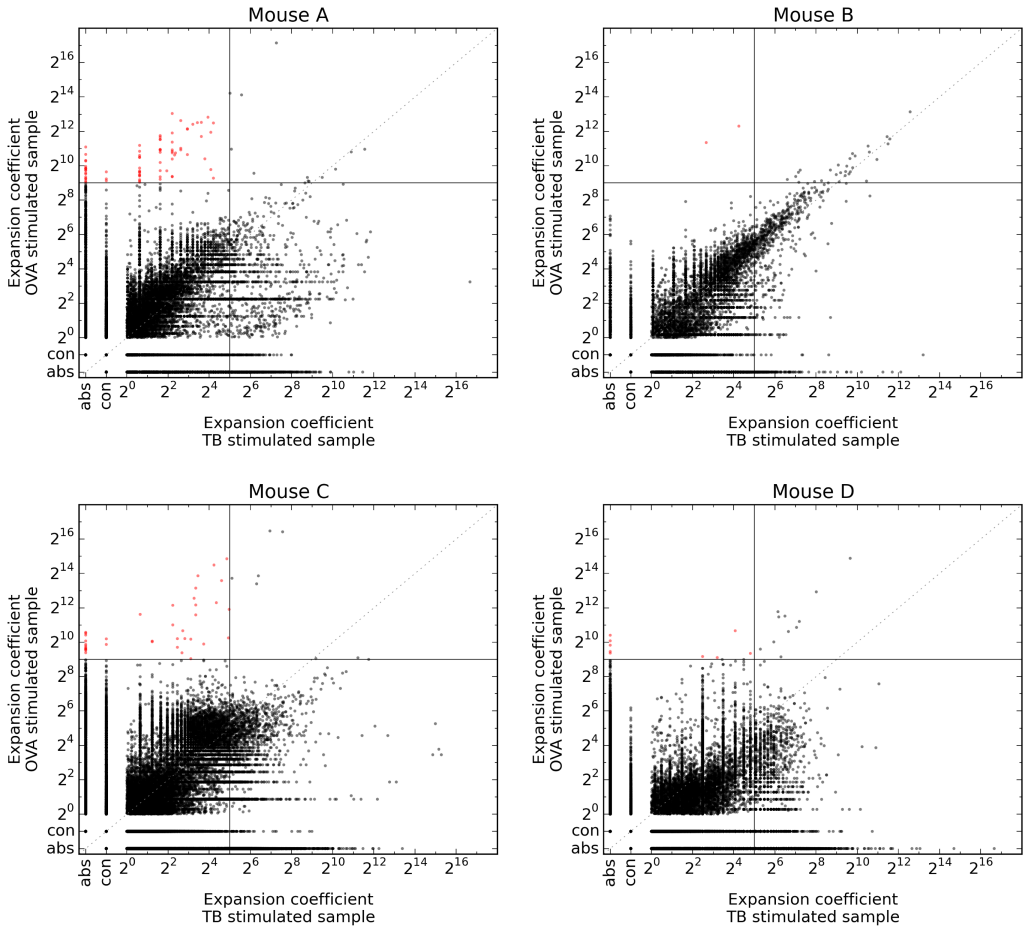

Figure 4.11: Correlation between CDR3β expansion coefficients in TB and OVA stimulated samples

For each CDR3β in an *in vitro* stimulated sample an expansion coefficient is calculated by dividing the frequency of the CDR3β by its frequency in the *ex vivo* spleen sample from the same animal. The relationship between the expansion coefficients for each CDR3β in TB and OVA *in vitro* stimulated sample. 'abs' denotes absence from the indicated sample and 'con' denotes contraction (i.e. expansion coefficient < 1). Diagonal line is $y = x$, horizontal and vertical lines indicate the expansion coefficients used as thresholds to define OVA-associated CDR3βs. OVA-associated CDR3βs are coloured red.

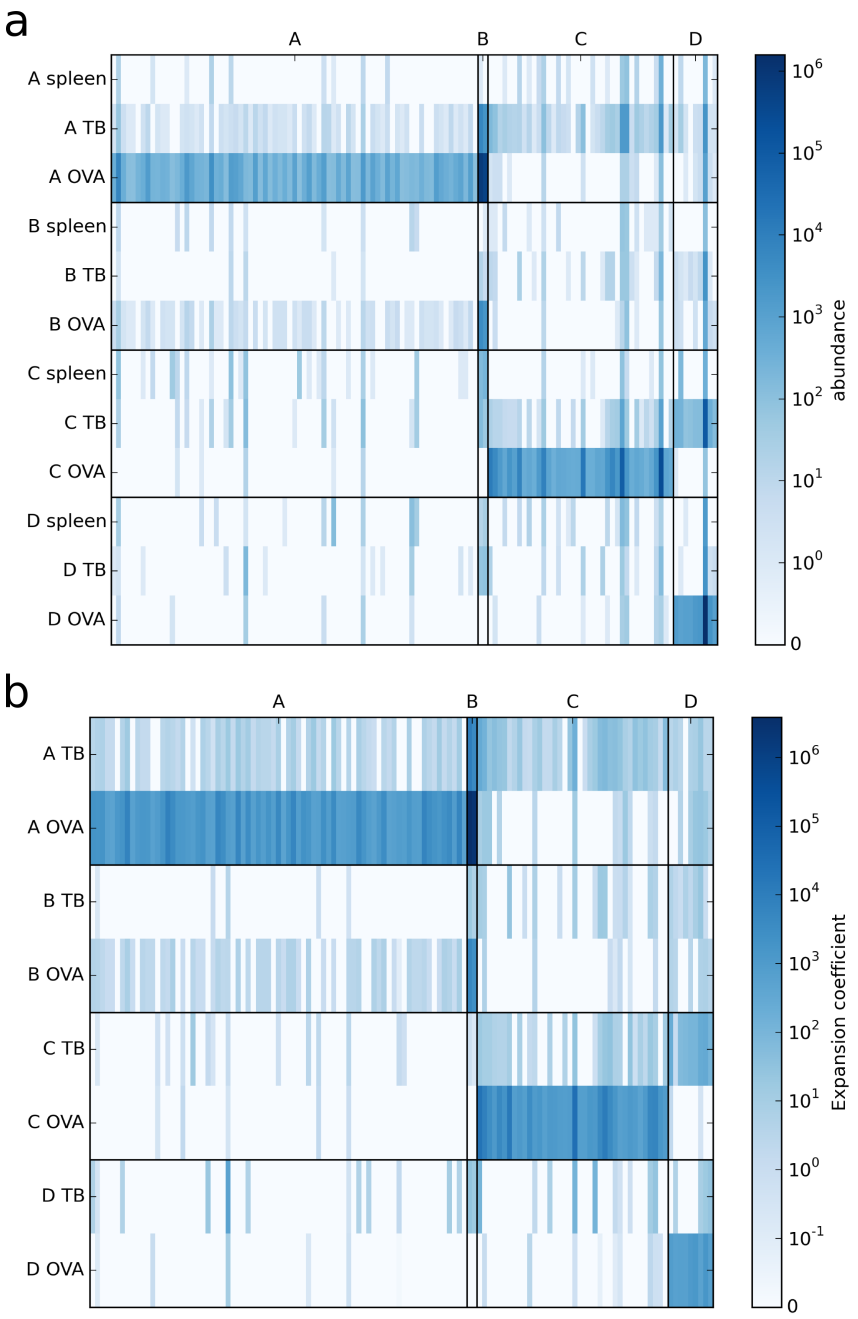

Figure 4.12: Patterns of abundance and expansion of OVA-associated CDR3βs
For the CDR3βs defined as OVA-associated in any of the four mice (124 in total), (a) their abundance in each of the samples from each of mice A-D, (b) their expansion coefficients (defined as frequency in stimulated sample divided by frequency in *ex vivo* spleen sample) in the TB and OVA stimulated samples from each mouse. CDR3βs are grouped by the mouse in which they are defined as OVA-associated.

grouped by the mouse in which they are defined as OVA-associated, in all of the samples considered here. This demonstrates that many of these CDR3βs are present in samples from other mice, but not at appropriate abundances to be considered OVA-associated. Figure 4.12b is an equivalent plot, but demonstrating the expansion coefficient of each OVA-associated CDR3β in each of the *in vitro* stimulated samples. These data suggest that in the *in vitro* stimulated samples there is differential selection of particular CDR3βs, such that some are able to be increased in abundance on OVA stimulation in samples from one mouse but not another.

Figure 4.12 also highlights some other interesting patterns of CDR3β response to *in vitro* OVA stimulation. The CDR3βs defined as OVA-associated in mouse A have a qualitatively similar pattern of abundance (and of expansion) in the samples from mice A and B (the CFA-only immunised mice), but do not demonstrate this pattern in mice C or D (the CFA+OVA immunised mice). This might suggest that these CDR3βs are only able to respond to *in vitro* OVA stimulation when there has been no pre-exposure (by *in vivo* immunisation) to OVA.

There are a number of CDR3βs (including the two defined as OVA-associated in mouse B) that demonstrate an opposite pattern of behaviour in the CFA-only immunised mice (mice A and B) to the CFA+OVA immunised mice (mice C and D). These CDR3βs are present at low abundance in the spleen samples, are somewhat expanded in TB stimulated samples and are more highly expanded in OVA stimulated samples from CFA-only immunised mice (A and B). In contrast, they are present at relatively high abundance in the spleen samples from CFA+OVA immunised mice (C and D), perhaps suggesting that they have been expanded *in vivo* in response to OVA. However, they are contracted or deleted in the OVA stimulated sample from the these mice, suggesting that *in vitro* stimulation with OVA is not able to cause proliferation of these CDR3βs when they have already been exposed to ovalbumin via *in vivo* immunisation.

Interestingly, the OVA-associated CDR3βs defined by their expansion coefficients in mouse D (CFA+OVA immunised) demonstrate completely opposite behaviour in mouse C (also CFA+OVA immunised), with high abundance and expansion coefficient after TB stimulation but not after OVA stimulation. There are also some CDR3βs

that are highly public, being abundant in all samples from all mice, but that are only responding substantially to OVA in one of the mice, again suggesting differential selection by OVA between these samples.

4.7 *In vitro* data: discussion

The TCR repertoires from *in vitro* stimulated samples of $CD4^+$ T cells have the potential to provide a clearer picture of the T cell response to OVA peptide, since repertoires can be measured from the same mice before and after *in vitro* culture. As seen in the *in vivo* data, we observed that the frequency of the CDR3βs from known OVA-responsive TCRs does not increase on OVA exposure (Figure 4.8) and that *in vitro* stimulation with the same antigen does not drive an increased similarity in repertoires (Figure 4.9).

An expansion coefficient can be calculated for each CDR3β in an *in vitro* stimulated sample with reference to the frequency of the CDR3β in the *ex vivo* sequenced spleen sample from the same mouse. This expansion coefficient provides a more direct estimate of the proliferation of a T cell clone in response to *in vitro* stimulation than was possible with the *in vivo* data. It is important to note that the expansion coefficient measured in this way captures the proliferation of a clone due to *in vitro* stimulation and does not include the effect of the *in vivo* immunisation.

For the four mice from which we have *ex vivo* spleen, TB stimulated and OVA stimulated samples, the number of CDR3βs that have an expansion coefficient above a given threshold in response to OVA stimulation does not appear to be dependent on previous OVA immunisation (Figure 4.10e). This suggests, unexpectedly, that the 'magnitude' of response to OVA stimulation is equivalent regardless of prior exposure to the antigen but additional data would be required to further test this finding. It might be interesting to test the hypothesis that the response to *in vitro* OVA stimulation by cells from CFA only mice is less functionally active despite being of the same 'magnitude' by obtaining functional read-outs from the samples, perhaps measuring T cell proliferation by CFSE dilution or effector function by IFNγ assays.

When we considered the relationship between expansion coefficients in TB stimulated

and OVA stimulated samples from the same mouse (Figure 4.11), we found that clones were equivalently expanded (statistically significant positive correlation) in both samples when the mouse had been immunised with CFA only, but were differentially expanded (a non-significant positive correlation) in the two samples when the mouse had been immunised with CFA+OVA. This suggests that in OVA containing wells without prior OVA exposure, this response is not specific since the same clones expand in response to TB stimulation. In contrast, the expansion in response to OVA peptide is not related to the expansion in response to TB stimulation when the mouse has been previously exposed to ovalbumin in the immunisation.

Comparison of expansion coefficients in OVA and TB stimulated samples from each mouse allowed us to define OVA-associated CDR3βs in a similar, if more direct, method as was used for the *in vivo* data. The identified CDR3βs are only defined as OVA-associated in a sample from one of the mice analysed. However, in contrast to our findings in the *in vivo* data, these CDR3βs are found to be relatively public in terms of their presence in other mice (Figure 4.12). In particular, some CDR3βs have high abundance (and high expansion coefficient) in the OVA stimulated sample, but not the TB stimulated sample from one mouse and exhibit opposite behaviour in another mouse. These data are highly suggestive of a differential and complicated pattern of selection of clones on *in vitro* stimulation in samples of cells from different animals. To confirm this, further experiments will be required to validate that the CDR3βs identified in this analysis are responding to OVA peptide (perhaps by tetramer sorting). It might also be interesting to test whether clones that are exhibiting differential behaviour in samples from different mice are functionally or transcriptionally distinct.

4.8 Identifying amino acid based motifs of OVA-associated CDR3βs

From the *in vivo* data (Section 4.4.3) and *in vitro* data (Section 4.6.3) three sets of CDR3βs have been identified that appear to be expanded in response to ovalbumin and are being referred to as 'OVA-associated' here. These sets are:

Set 1 CDR3βs that are highly expanded *in vivo* (expansion coefficient $> 2^9$) relative to unimmunised samples in at least one CFA+OVA immunised spleen sample and are not highly expanded (expansion coefficient $< 2^4$) in any of the CFA only immunised spleen samples.

Set 2 CDR3βs that are highly expanded *in vivo* (expansion coefficient $> 2^9$) relative to unimmunised samples in at least one CFA+OVA immunised spleen sample and are not classified as stimulated by endogenous peptide or CFA (i.e. maximum frequency in either CFA or unimmunised samples is at least 2 fold smaller than minimum non-zero frequency in CFA+OVA samples).

Set 3 CDR3βs that are highly expanded (expansion coefficient $> 2^9$) relative to spleen sample from the same mouse in an *in vitro* OVA stimulated sample but are not highly expanded (expansion coefficient $< 2^5$) in the TB stimulated sample from the same mouse.

Set 1 contains 603 CDR3βs, of which 160 are also present in set 2 (Figure 4.13a). Set 2 contains 245 CDR3βs in total, and set 3 consists of 124 CDR3βs, none of which are present in sets 1 or 2.

To assess whether there is an amino acid sequence motif which characterises OVA associated CDR3βs the pairwise distance or similarity between CDR3βs within each of the sets is measured. We used the Levenshtein distance (the number of insertions, substitutions or deletions required to transform one string into another) as a distance metric and the p-spectrum kernel (number of substrings of length p shared between two strings) as a similarity metric. For comparison, we repeatedly randomly selected sets of CDR3βs from the total combined data, with these control sets being size-matched to the OVA-associated sets. We measured pairwise distance and similarity in the control sets in the same way as for the OVA-associated sets. In order to reduce any contribution of sequencing error to the control sets, only CDR3βs which appeared at least 5 times across the data were available for sampling.

The CDR3βs from the *in vitro* (Set 3), but not the *in vivo* (Set 1 and Set 2) sets of OVA-associated CDR3βs, showed a reduced average pairwise Levenshtein distance,

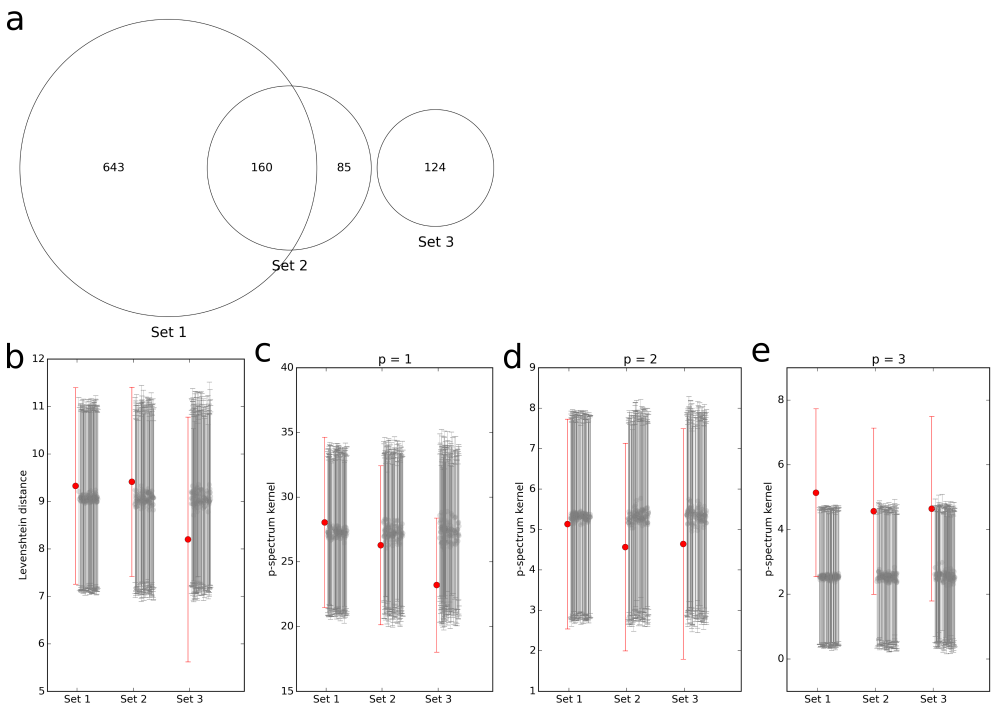

Figure 4.13: Sets of OVA-associated CDR3βs demonstrate amino acid sequence similarity For the three sets of CDR3βs identified in sections 4.4.3 and 4.6.3: (a) overlap between the sets. Mean ± standard deviation of (b) pairwise Levenshtein distance between CDR3βs within a set (red) and (c-e) pairwise similarity between CDR3βs within a set (red), measured by p-spectrum kernels, compared to the same measure in size-matched random samples of CDR3βs (grey).

compared to random sampling (Figure 4.13b). However, increased similarity could be observed for both *in vitro* and *in vivo* OVA associated CDR3β sets when using p-spectrum kernels with $p = 3$ (Figure 4.13e). Shorter length spectrum kernels did not show any increased similarity (Figures 4.13c and d).

The results presented in Figure 4.13 suggest that short contiguous stretches of amino acids might define enriched motifs within the sets of OVA-associated CDR3βs. A similar phenomenon was observed when the usage of each p-tuplet, for $p = 1$,2 and 3 (i.e. amino acid, amino acid pair and amino acid triplet) in the OVA-associated sets was counted, and ranked in comparison to the usage counts in 500 random CDR3β samples (selected as detailed above). Results are expressed as the percentile rank of the usage found in OVA associated sets in comparison to random sets.

No individual amino acid is consistently over- or under-represented across all three sets of OVA associated CDR3βs (Figure 4.14a). However seven pairs of amino acids and

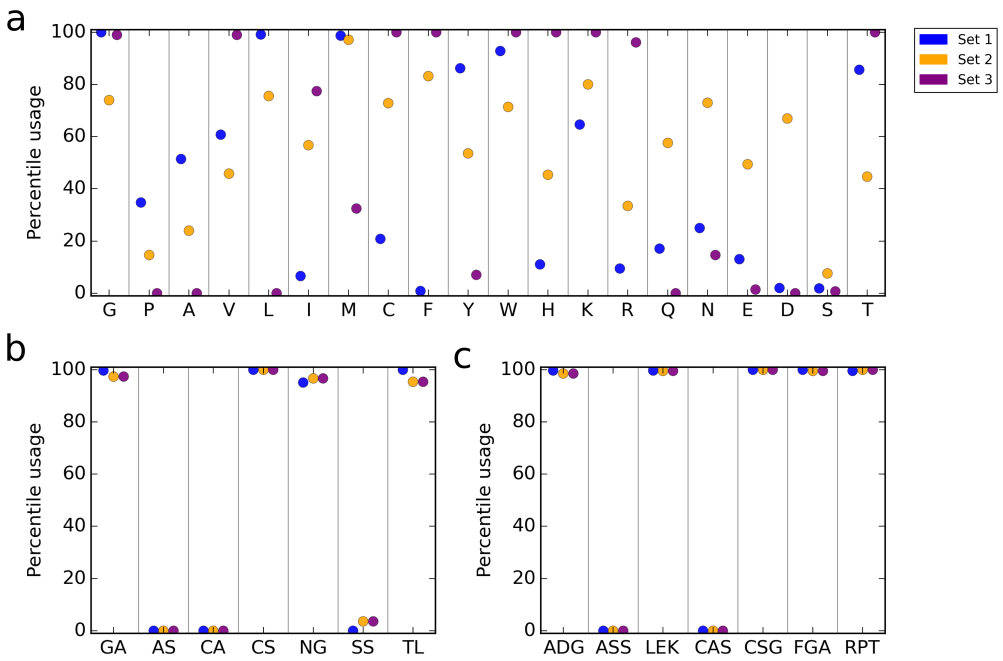

Figure 4.14: Amino acid motif usage in OVA-associated sets of CDR3βs
For the three sets of CDR3βs identified in sections 4.4.3 and 4.6.3:
(a) The amino acid usage within the sets of CDR3βs, expressed as the percentile rank when compared to the usage observed in repeated random size-matched samples.
(b) and (c) The usage within the OVA-associated sets, expressed as percentile rank as in (a), of amino acid pairs and triplets which have high or low expression in all three sets.

seven triplets are over- or under-represented ($> 95^{\text{th}}$ or $< 5^{\text{th}}$ percentile of usage in the control randomly sampled sets) in all three sets (Figure 4.14b,c). Many of these amino acid motifs appear to be related to the V or J encoded parts of the CDR3β (e.g. CA, CS, CAS, FGA), but others, such as LEK and RPT, likely occur towards the center of the CDR3β and could be responsible for contact with OVA peptide.

The presence of these amino acid duplet and triplet motifs in the full sequenced repertoires in spleen samples from immunised and unimmunised mice is counted (Figure 4.15a and b). None of the motifs alone distinguish CFA only from CFA+OVA immunised samples, although the usage of AS, CA, ASS and CAS is significantly different (KS test, $p < 0.01$) between unimmunised and immunised samples. Neither does the presence of a combination of motifs within a CDR3β differentiate between CFA only and CFA+OVA immunised samples (Figure 4.15c and d) with the proportion of CDR3βs that contain at least x of the over-expressed duplet or triplet motifs not

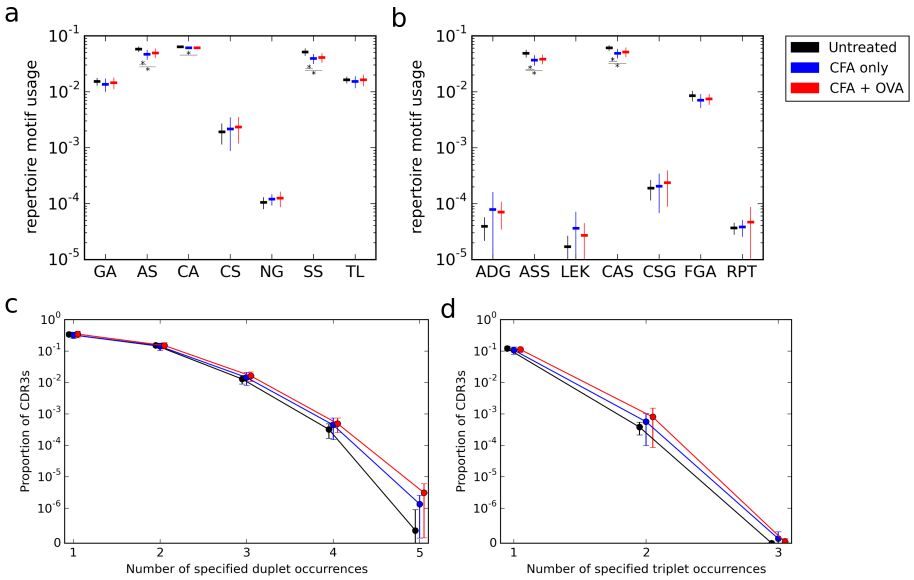

Figure 4.15: Usage of OVA associated duplets and triplets in full CDR3β repertoires
(a) and (b) The proportional usage (mean $\pm$ standard deviation) in the full repertoire samples of *in vivo* samples from immunised mice of amino acid duplets and triplets identified as being consistently over- or under-expressed in sets of OVA-associated CDR3βs.
(c) and (d) The proportion of CDR3βs in an *in vivo* sample (mean $\pm$ standard deviation) that contain multiple (at least the number shown on the x-axis) duplets or triplets that are over-expressed in OVA-associated CDR3βs.
Immunisation status of the mouse is indicated by colour for all panels.

being statistically different between CFA only and CFA+OVA immunised samples for any value of x.

4.9 Discussion

Identification of a TCR signature of immune response to a single antigen is an essential initial step in exploiting the potential of the TCR repertoire to act as a marker or signature of the immune challenges faced by an individual. The characteristics of such a signature may also reveal interesting features of the T cell response, both in terms of the structural recognition of antigen by individual TCRs and at a T cell population level (breadth, diversity etc.). In this chapter, the TCR repertoire of genetically identical mice in response to immunisation with ovalbumin has been considered, to attempt to identify a signature of antigen-specific T cell response above a background of response to an adjuvant. In the data analysed in this chapter the adjuvant employed in the experimental model is CFA, commonly used to induce a response to ovalbumin e.g.

[13, 43, 73].

We first looked for CDR3βs that are part of previously reported OVA-specific TCRs. Unexpectedly, the CDR3βs of these known OVA-specific clones are present at a similar frequency in samples from CFA+OVA immunised mice as in CFA immunised mice; and similar frequency after *in vitro* stimulation with OVA or with TB. These results suggest either that these 'public' β chains are found paired with many different αs, giving differing TCR specificities only some of which are OVA specific, while others are specific for some component of MTb; or that the reported OVA specific clones using these β chains are in fact cross-reactive to peptides from the adjuvant as well as from ovalbumin. Using the frequency of previously reported antigen specific clones does not appear to be a feasible approach to defining a signature of antigen exposure in these data.

The approach used in this work to identify OVA-associated CDR3βs has been to estimate an 'expansion coefficient' for each CDR3β in an immunised or stimulated sample, calculated as its observed frequency in the sample divided by the frequency in a 'control' sample. For the *in vivo* samples, the control is pooled unimmunised repertoires, while for the *in vitro* stimulated samples the control is the *ex vivo* spleen sample from the same mouse. Then OVA-associated CDR3βs can be defined on the basis of having a high expansion coefficient in an OVA exposed repertoire, but not in repertoires not exposed to OVA. The analysis in this chapter defines and analyses three sets of OVA-associated CDR3βs: two from *in vivo* data, using different criteria to exclude CDR3βs, and one set from *in vitro* data.

The key feature of all three sets of OVA-associated CDR3βs is that they are almost all private sequences. The majority of CDR3βs defined as OVA-associated only have this property in one of the OVA immunised mice or one of the OVA stimulated *in vitro* samples, suggesting a highly individual specific clonal response. This is also consistent with the lack of increased overlap between repertoires after immunisation with ovalbumin, compared to unimmunised pairs: ovalbumin immunisation does not drive a convergent CDR3β repertoire. Finding a single CDR3β, or even a small set of CDR3βs, whose abundance in a sample can predict antigen exposure also does not

therefore appear to be a feasible approach, even in this relatively simple experimental model.

The privacy of OVA-associated CDR3βs from the *in vivo* repertoires appears to be a feature of availability: the CDR3β that are OVA-associated in one mouse are often simply not present in the sampled repertoires of the majority of the other mice (Figure 4.6h). The simplest interpretation is therefore that the privacy of the OVA-associated CDR3βs is due to effects of stochastic repertoire generation rather than differential selection after immunisation. However, in the *in vitro* stimulated data the CDR3βs that are defined as OVA-associated in one mouse are often present, but not defined as OVA-associated, in other mice (Figure 4.12), suggestive of differential antigen selection in different animals. However, without paired $\alpha\beta$ sequence data we cannot conclusively assess whether selection of T cell clones to proliferate in response to the *in vitro* antigen stimulation is differential in these repertoires, or whether *in vitro* expansion favours more 'public' T cell clones.

The overwhelming privacy of the OVA-associated CDR3βs suggests that any signature of ovalbumin exposure will require a metric which will map 'similar' TCRs by sequence to sets of similar TCRs by antigen recognition. A number of detailed studies have identified particular residues of the peptide bound within the MHC binding pocket which are key for TCR recognition [19, 24]. We hypothesised that these specific antigen residues may impose some sequence restrictions on the CDR3 of the cognate TCR. We therefore further hypothesised that the set of OVA peptide specific TCRs might be enriched for common short stretches of amino acids necessary for this specific antigen recognition. We initially explored this hypothesis by measuring the amino acid sequence-level similarity of the OVA-associated sets of CDR3s using different measures of sequence distance or similarity. We observed that, as predicted, string kernel measures (which capture the number of shared short amino acid stretches between two CDR3βs) showed that these sets are more similar than randomly selected control sets when triplets of amino acids are considered. Consistent with this observation, we also found that some specific amino acid triplets were found to be either over- or under-represented in OVA-associated sets of CDR3βs compared to randomly sampled sets of CDR3βs. Interestingly, a number of these motifs appear to be present in germline V

or J regions, and hence found towards the beginning or end of the CDR3 region. In contrast, others were not germline encoded and look to be present in the non-templated middle of the CDR3. The frequency of each over- or under-represented motif alone, however, was not sufficient to discriminate between OVA and CFA+OVA repertoires. Current studies in our group are exploring how machine learning techniques can be used to select amino acid motif 'features' from repertoires and combine them to produce classifiers to predict the immunisation status from these TCR repertoires.

Limitations of the data and further work

The ability of this study to identify a TCR signature of response to ovalbumin is limited in a number of respects. Key to the work is a successful immunisation of the mice, particularly with respect to the ovalbumin. The inclusion of functional assays, such as measurement of IFNγ production or Ki67 expression of the isolated CD4$^+$ T cells that are taken for sequencing, could have confirmed an increase in activation of T cells on immunisation and might be considered for future experiments. The success of the ovalbumin immunisation could be confirmed by measuring the proportion of T cells from each mouse that are positive in an OVA peptide-MHC multimer assay, and additionally multimer sorted cells could be taken for sequencing to validate the CDR3βs defined as OVA-associated in each mouse using the techniques described in this chapter.

The analysis in this chapter is heavily reliant on measurements of the abundance of each CDR3β in each sample. As discussed in Chapter 2, there is variability in PCR amplification efficiencies which might affect these results. Additional experiments have been performed by Nir Friedman's group (currently being sequenced) which use a unique molecular barcoding protocol, allowing more reliable quantification of CDR3β abundance.

The *ex vivo* spleen samples taken at early, mid and late timepoints after immunisation potentially allow the kinetics of the response to ovalbumin immunisation to be studied. The sample sizes in these experiments, with only a few mice at each time point, restricts our ability to draw general conclusions from the repertoires and additional samples from each timepoint would allow more detailed analysis. Alternatively, longitudinal

samples after immunisation, perhaps in the form of repeated blood draws, might give smaller but more informative TCR repertoire samples.

The *in vitro* stimulation data could provide a cleaner system than *in vivo* immunisation to identify motifs of TCR repsonse to peptides. A more comprehensive set of samples, including T cells from unimmunised mice and from mice at different timepoints after immunisation, could allow for a more robust signature identification. The immunised mice in the *in vitro* stimulation experiments used in this chapter were sacrificed at 7 days post immunisation, probably at the peak of effector T cell response. T cells from mice after the acute immune response has resolved and T cell memory has been established might demonstrate a response on *in vitro* stimulation with a clearer ovalbumin signature.

It might be useful if the sequence data was accompanied by phenotypic information regarding the sequenced cells. Splitting T cells into subsets (e.g. naive, effector, memory) before sequencing could help with identification of the clones that are responsible for response to ovalbumin. Obtaining paired TCR$\alpha\beta$ information, either using single cell analysis, protocols which allow paired sequence analysis or possibly by post-sequencing computational techniques, would also go some way to strengthening the conclusions summarised above and answering outstanding questions regarding the results seen in this work.

Chapter 5

Discussion

The TCR repertoire is inherently complicated, containing a large number of different clones, each present at an abundance which changes both in response to signals from the environment and in relation to the other clones present. Therefore the TCR repertoire is difficult to measure and analyse, but also has the potential to provide much information regarding the status of an individual's immune system. In this thesis a number of aspects of the TCR repertoire have been considered. Much of the work presented attempts to capture the complexity of data in the form of models describing the underlying behaviour of a system.

In Chapter 2, we observed that the efficiency of PCR amplification was very heterogenous, even when reaction conditions and nucleotide sequences are kept as constant as possible. We model PCR amplification as modified branching processes, and find that the observed 'family sizes' after amplification cannot be explained with any intuitive process, but require inheritance of variable amplification efficiencies from one molecule to its copies. The data described in this chapter demonstrate that for sequencing studies requiring robust quantification of multiple types of initial molecules, such as immune repertoire studies, unique molecular barcoding prior to PCR amplification is essential.

The model of maintenance of self-tolerance presented in Chapter 3 describes a way in which the output of the stochastic TCR rearrangement process can be reshaped in order to produce a T cell population which is tolerant to the multitude of self-antigens

continually presented by DC. This is achieved via a relatively simple model, which can be expressed as a series of linear inequalities, involving DC integrating TCR binding signals received, and acting to adjust T cell clone abundances when too much signal is experienced. This model restricts total T cell affinity to self-antigens, while preserving the known features of the TCR repertoire, including clone richness, heterogenous clone sizes and cross-reactivity between clones and peptides.

Sequenced TCR repertoires from immunised mice are analysed in Chapter 4, demonstrating the complexity present in the immune response even in a relatively simple model system. We find extraordinarily private CDR3β responses to ovalbumin in these mice, including evidence of differential selection of clones in samples taken from different mice and stimulated in vitro. However, there is some evidence that the clones associated with the response to ovalbumin in these data share features at the amino acid sequence level. We show that studying the CDR3βs present in the repertoire does not appear to identify cross-individual signatures of immune response against a particular antigen, likely due to both the privacy of the clones present and their cross-reactivity. However, this work suggests that searching for motifs of amino acids defining antigen response might be more fruitful.

Limitations and further work

Amplification heterogeneity in the PCR reaction

The data and models presented in Chapter 2 demonstrate that, in order for PCR amplification to result in the heterogenous family sizes observed, some form of inheritance, from one molecule to its copies present in the next cycle, of a variable amplification efficiency is required. When these molecules are identical for the majority of their nucleotide sequence (apart from the molecular barcode) and conditions within the PCR reaction should ensure the samples are well mixed it is hard to envisage what mechanism could explain this inheritance. Additional experiments to confirm that the barcode is not causing the differential amplification observed would be useful in strengthening the results of the work, as would a clearer idea of the molecular mechanism that could cause the observed heterogeneity. More detailed modelling of the PCR process itself,

rather than simply modelling stochastic replication events, might help in this respect.

Model of maintenance of immune tolerance in the periphery

The work presented in this thesis on the model of maintenance of immune tolerance (Chapter 3) is by necessity a simplified model of potential reshaping of the T cell population through interactions between T cells and resting DC. A major assumption of the current implementation of the model is that within the time period that a DC is able to integrate signal over, it experiences enough T cell/DC interactions such that the switch to a tolerising state is essentially based on the number of cognate T cells in the whole population. Similarly, we assume that 'kill' signals from a tolerising DC affect all T cells when averaged over a sufficient number of time steps.

This assumption can be addressed by using an agent based model (ABM) implementation, where each DC and each T cell is modelled explicitly, rather than modelling T cell clones and self-peptide presentation profiles, as is presented in this thesis. Random encounters between T cells and DCs would result in signal integration on the part of the DC in order to determine whether to switch to a tolerising state, as well as signal integration of 'survival' or 'death' signals on the part of the T cells.

An ABM implementation would also enable another limitation of the work presented here to be addressed more easily. The results presented in Chapter 3 are all based on the state of the T cell population while DCs are all resting. It does not consider activation of the DCs by signals from other innate immune system cells, and therefore does not explicitly model a T cell immune response. An ABM would allow exploration of the situation where some DCs begin to present non-self (or previously unencountered) peptides, along with activation signals, to T cells of a self-tolerant population. Questions such as how well self-tolerance can be maintained by the DCs that remain resting while others are activated could be asked with such a model, allowing predictions to be made regarding the extent of immune activation before autoimmunity becomes harmful.

The model implemented in Chapter 3 makes a number of parametrisation assumptions, particularly around the proliferation rate of T cells, the rate at which tolerising DCs are able to kill T cells, and the distribution of binding strengths between TCRs and peptide-

MHC complexes. These could be improved with more experimental data. For instance, screening a range of TCR clones for affinity to a range of peptide-MHC tetramers could provide information on the sparsity of the matrix Q describing the TCR-pMHC affinities as well as the distribution of its non-zero values. Additionally, the current implementation of the model is not operating at physiological scales of number of T cells and APCs. Parallelisation and use of computing clusters would allow more physiologically relevant simulations to be performed and therefore more accurate predictions to be made.

Private T cell responses to ovalbumin

The mouse TCR repertoire sequence data analysed in Chapter 4 has a number of limitations. Crucially, it is not barcoded data which means conclusions drawn on the basis of clone abundance within samples are less robust. A repeated experiment, using a protocol that uniquely labels each molecule of TCR RNA before amplification, has been performed and the data should be available soon. Analysis of the barcoded data should allow for confirmation or rejection of the finding that the CDR3βs that are highly abundant after immunisation with ovalbumin are private to each animal.

Additionally, the analysis of these repertoire data are limited by only including β chains from each TCR. We are unable to say anything about T cell clones that might be responding to ovalbumin, but only about the TCRβs, which might each be paired with multiple αs. Extraction and sequencing of the corresponding α chains, preferably paired with the β occuring in the same heterodimer, would greatly increase the specificity of the conclusions that can be drawn.

If additional experiments could be performed, it would be useful to immunise with the same model antigen in the context of different adjuvants. If particular amino acid motifs comprise a signature of T cell response to ovalbumin this should be the case in multiple contexts, and a signature derived from CFA+OVA data should be able to distinguish OVA immunised repertoires in conjunction with other adjuvants.

General considerations

The work presented in this thesis uses modelling and computational techniques to extract useful information from what could be termed 'big data'. As experimental techniques develop and the data produced becomes 'larger', both in terms of the number of observations measured and the dimensionality of the measurements for each observation, analysis becomes more complicated. In the majority of realistic experimental models, multi-dimensional measurements mean that system-level effects become very apparent. A single perturbation to an experimental system results in multiple changes in 'outputs' measured, and in order to obtain a desired change in the measurements in the model there might be multiple possible interventions. These observations, and the combinatorial effect of multiple possible interventions, means that mathematical modelling and other computational approaches are essential to gain understanding of the behaviour of an experimental system and to direct experimental effort.

A common feature emerging from analysis of multi-dimensional and high throughput data from many experimental models is that individual cells often demonstrate completely different behaviours in response to the same conditions, which would not be revealed in measurements on bulk populations. However, it is often the bulk population measurement that is the objective of experimental or medical intervention, and that is thought to be clinically relevant. Mathematical models, parametrised by data from single cell experiments, will be invaluable in these situations to understand the effect of single cell behaviour on bulk population measurements and to predict what level of intervention is required to reach a desired effect.

In terms of the TCR repertoire, acquistion of large datasets appears to be progressing faster than analysis frameworks are developing. Any application of TCR whole repertoire sequencing and analysis to address clinical questions regarding diagnosis or prediction of prognosis requires assessment of patient repertoire in comparison to healthy controls. However, although common features across all healthy repertoires can be identified (for instance, relatively consistent gene usage, clone size distributions and diversity metrics) it is unclear which characteristics of the repertoire are essential for 'health'. Similarly it is not obvious how changes to any of the measured features of

the TCR repertoire affect the individual's ability to respond to immune challenge or to clinical intervention in disease. It is essential that the variability in repertoires amongst healthy individuals is well understood before we are able to apply and interpret measures of repertoire in clinical settings.

Bibliography

[1] John Aach and George M. Church. Mathematical models of diffusion-constrained polymerase chain reactions: basis of high-throughput nucleic acid assays and simple self-organizing systems. *Journal of Theoretical Biology*, 228(1):31–46, 2004.

[2] Kaveh Abdi, Nevil J. Singh, and Polly Matzinger. Lipopolysaccharide-activated dendritic cells: "exhausted" or alert and waiting? *The Journal of Immunology*, 188(12):5981–5989, 2012.

[3] Alfred V. Aho and Margaret J. Corasick. Efficient string matching: an aid to bibliographic search. *Communications of the ACM*, 18(6):333–340, 1975.

[4] Daniel Aird, Michael G. Ross, Wei-Sheng Chen, Maxwell Danielsson, Timothy Fennell, Carsten Russ, David B. Jaffe, Chad Nusbaum, and Andreas Gnirke. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, 12(2):R18, 2011.

[5] Eltaf Alamyar, Véronique Giudicelli, Shuo Li, Patrice Duroux, and Marie Paule Lefranc. IMGT/Highv-quest: The IMGT web portal for immunoglobulin (IG) or antibody and T cell receptor (TR) analysis from NGS high throughput and deep sequencing. *Immunome Research*, 8(1):1–15, 2012.

[6] H. K. Alexander and L. M. Wahl. Self-tolerance and Autoimmunity in a Regulatory T Cell Model. *Bulletin of Mathematical Biology*, 73(1):33–71, 2011.

[7] Shahar Alon, Francois Vigneault, Seda Eminaga, Danos C. Christodoulou, Jonathan G. Seidman, George M. Church, and Eli Eisenberg. Barcoding bias in high-throughput multiplex sequencing of miRNA. *Genome Research*, 21(9):1506–1511, 2011.

[8] F W Alt, E M Oltz, F Young, J Gorman, G Taccioli, and J Chen. VDJ recombination. *Immunology today*, 13(8):306–314, 1992.

[9] Mark S. Anderson, Emily S. Venanzi, Ludger Klein, Zhibin Chen, Stuart P. Berzins, Shannon J. Turley, Harald von Boehmer, Roderick Bronson, Andree Dierich, Christophe Benoist, and Diane Mathis. Projection of an Immunological Self Shadow Within the Thymus by the Aire Protein. *Science*, 298(5597):1395–1401, 2002.

[10] Sanjeev Arora, Elad Hazan, and Satyen Kale. The Multiplicative Weights Update Method: a Meta-Algorithm and Applications. *Theory of Computing*, 8(1):121–164, 2012.

[11] T. P. Arstila, Armanda Casrouge, Veronique Baron, Jos Even, Jean Kanellopoulos, and Philippe Kourilsky. A Direct Estimate of the Human T Cell Receptor Diversity. *Science*, 286(5441):958–961, 1999.

[12] Iren Bains, Hisse M. van Santen, Benedict Seddon, and Andrew J. Yates. Models of Self-Peptide Sampling by Developing T Cells Identify Candidate Mechanisms of Thymic Selection. *PLoS Computational Biology*, 9(7):e1003102, 2013.

[13] Lucinda Beck and Hans L. Spiegelberg. The polyclonal and antigen-specific IgE and IgG subclass response of mice injected with ovalbumin in alum or complete Freund's adjuvant. *Cellular Immunology*, 123(1):1–8, 1989.

[14] Joost B. Beltman, Athanasius F.M. Marée, Jennifer N. Lynch, Mark J. Miller, and Rob J. de Boer. Lymph node topology dictates T cell migration behavior. *The Journal of Experimental Medicine*, 204(4):771–780, 2007.

[15] David R. Bentley, Shankar Balasubramanian, Harold P. Swerdlow, Geoffrey P. Smith, John Milton, Clive G. Brown, Kevin P. Hall, Dirk J. Evers, Colin L. Barnes, Helen R. Bignell, Jonathan M. Boutell, Jason Bryant, Richard J. Carter, R. Keira Cheetham, Anthony J. Cox, Darren J. Ellis, Michael R. Flatbush, Niall Gormley, Sean J. Humphray, Leslie J. Irving, Mirian S. Karbelashvili, Scott M. Kirk, Heng Li, Xiaohai Liu, Klaus S. Maisinger, Lisa J. Murray, Bojan Obradovic, Tobias Ost, Michael L. Parkinson, Mark R. Pratt, Isabelle M. J. Rasolonjatovo, Mark T. Reed, Roberto Rigatti, Chiara Rodighiero, Mark T. Ross, Andrea Sabot, Subramanian V. Sankar, Aylwyn Scally, Gary P. Schroth, Mark E. Smith, Vincent P. Smith, Anastassia Spiridou, Peta E. Torrance, Svilen S. Tzonev, Eric H. Vermaas, Klaudia Walter, Xiaolin Wu, Lu Zhang, Mohammed D Alam, Carole Anastasi, Ify C Aniebo, David M D Bailey, Iain R Bancarz, Saibal Banerjee, Selena G Barbour, Primo Baybayan, Vincent Benoit, Kevin F Benson, Claire Bevis, Phillip J Black, Asha Boodhun, Joe S Brennan, John Bridgham, Rob C. Brown, Andrew Brown, Dale H Buermann, Abass Bundu, James C Burrows, Nigel P Carter, Nestor Castillo, Maria Chiara E. Catenazzi, Simon Chang, R Neil Cooley, Natasha R Crake, Olubunmi O Dada, Konstantinos D Diakoumakos, Belen Dominguez-Fernandez, David J Earnshaw, Ugonna C Egbujor, David W Elmore, Sergey S Etchin, Mark R Ewan, Milan Fedurco, Louise J Fraser, Karin V Fuentes Fajardo, W Scott Furey, David George, Kimberley J Gietzen, Colin P Goddard, George S Golda, Philip Granieri, David E Green, David L Gustafson, Nancy F Hansen, Kevin Harnish, Christian D Haudenschild, Narinder I Heyer, Matthew M Hims, Johnny T Ho, Adrian M Horgan, Katya Hoschler, Steve Hurwitz, Denis V Ivanov, Maria Q Johnson, Terena James, T Huw Jones, Gyoung-Dong Kang, Tzvetana H Kerelska, Alan D Kersey, Irina Khrebtukova, Alex P Kindwall, Zoya Kingsbury, Paula I Kokko-Gonzales, Anil Kumar, Marc A Laurent, Cynthia T Lawley, Sarah E. Lee, Xavier Lee,

Arnold K Liao, Jennifer A Loch, Mitch Lok, Shujun Luo, Radhika M Mammen, John W Martin, Patrick G McCauley, Paul McNitt, Parul Mehta, Keith W Moon, Joe W Mullens, Taksina Newington, Zemin Ning, Bee Ling Ng, Sonia M Novo, Michael J. ONeill, Mark A Osborne, Andrew Osnowski, Omead Ostadan, Lambros L Paraschos, Lea Pickering, Andrew C Pike, Alger C Pike, D Chris Pinkard, Daniel P Pliskin, Joe Podhasky, Victor J Quijano, Come Raczy, Vicki H Rae, Stephen R Rawlings, Ana Chiva Rodriguez, Phyllida M Roe, John Rogers, Maria C Rogert Bacigalupo, Nikolai Romanov, Anthony Romieu, Rithy K Roth, Natalie J Rourke, Silke T Ruediger, Eli Rusman, Raquel M Sanches-Kuiper, Martin R Schenker, Josefina M Seoane, Richard J Shaw, Mitch K Shiver, Steven W Short, Ning L Sizto, Johannes P Sluis, Melanie A Smith, Jean Ernest Sohna Sohna, Eric J Spence, Kim Stevens, Neil Sutton, Lukasz Szajkowski, Carolyn L Tregidgo, Gerardo Turcatti, Stephanie VandeVondele, Yuli Verhovsky, Selene M Virk, Suzanne Wakelin, Gregory C Walcott, Jingwen Wang, Graham J Worsley, Juying Yan, Ling Yau, Mike Zuerlein, Jane Rogers, James C Mullikin, Matthew E Hurles, Nick J McCooke, John S West, Frank L Oaks, Peter L Lundberg, David Klenerman, Richard Durbin, and Anthony J Smith. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, 2008.

[16] A. M. Berkley and P. J. Fink. Cutting Edge: CD8+ Recent Thymic Emigrants Exhibit Increased Responses to Low-Affinity Ligands and Improved Access to Peripheral Sites of Inflammation. *The Journal of Immunology*, 193(7):3262–3266, 2014.

[17] Katharine Best, Benny Chain, and Chris Watkins. Immune Tolerance Maintained by Cooperative Interactions between T Cells and Antigen Presenting Cells Shapes a Diverse TCR Repertoire. *Frontiers in Immunology*, 6:360, 2015.

[18] Katharine Best, Theres Oakes, James M. Heather, John Shawe-Taylor, and Benny Chain. Computational analysis of stochastic heterogeneity in PCR amplification efficiency revealed by single molecule barcoding. *Scientific Reports*, 5:14629, 2015.

[19] Michael E. Birnbaum, Juan L. Mendoza, Dhruv K. Sethi, Shen Dong, Jacob Glanville, Jessica Dobbins, Engin Ozkan, Mark M. Davis, Kai W. Wucherpfennig, and K. Christopher Garcia. Deconstructing the peptide-MHC specificity of T cell recognition. *Cell*, 157(5):1073–87, 2014.

[20] P J Bjorkman, M a Saper, B Samraoui, W S Bennett, J L Strominger, and D C Wiley. The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens. *Nature*, 329(6139):512–518, 1987.

[21] Dmitriy A. Bolotin, Mikhail Shugay, Ilgar Z. Mamedov, Ekaterina V. Putintseva, Maria A. Turchaninova, Ivan V. Zvyagin, Olga V. Britanova, and Dmitriy M. Chudakov. MiTCR: software for T-cell receptor sequencing data analysis. *Nature Methods*, 10(9):813–814, 2013.

[22] Dmitry A. Bolotin, Ilgar Z. Mamedov, Olga V. Britanova, Ivan V. Zvyagin, Dmitriy Shagin, Svetlana V. Ustyugova, Maria A. Turchaninova, Sergey Lukyanov, Yury B. Lebedev, and Dmitriy M. Chudakov. Next generation sequencing for TCR repertoire profiling: Platform-specific features and correction algorithms. *European Journal of Immunology*, 42(11):3073–3083, 2012.

[23] Onur Boyman, Carsten Krieg, Dirk Homann, and Jonathan Sprent. Homeostatic maintenance of T cells and natural killer cells. *Cellular and Molecular Life Sciences*, 69(10):1597–1608, 2012.

[24] Robert D. Bremel and E. Jane Homan. Extensive T-Cell Epitope Repertoire Sharing among Human Proteome, Gastrointestinal Microbiome, and Pathogenic Bacteria: Implications for the Definition of Self. *Frontiers in Immunology*, 6(538), 2015.

[25] O. V. Britanova, E. V. Putintseva, M. Shugay, E. M. Merzlyak, M. A. Turchaninova, D. B. Staroverov, D. A. Bolotin, S. Lukyanov, E. A. Bogdanova, I. Z. Mamedov, Y. B. Lebedev, and D. M. Chudakov. Age-Related Decrease in TCR Repertoire Diversity Measured with Deep and Normalized Sequence Profiling. *The Journal of Immunology*, 192(6):2689–2698, 2014.

[26] Xavier Brochet, Marie Paule Lefranc, and Véronique Giudicelli. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. *Nucleic Acids Research*, 36:W503–W508, 2008.

[27] F. M. Burnet. A modification of Jerne's theory of antibody production using the concept of clonal selection. *CA: a cancer journal for clinicians*, 26(2):119–121, 1976.

[28] Thomas Charles Butler, Mehran Kardar, and Arup K. Chakraborty. Quorum sensing allows T cells to discriminate between self and nonself. *Proceedings of the National Academy of Sciences of the United States of America*, 110(29):11833–8, 2013.

[29] R. E. Callard, R. J. Armitage, W. C. Fanslow, and M. K. Spriggs. CD40 ligand and its role in X-linked hyper-IgM syndrome. *Immunology today*, 14(11):559–564, 1993.

[30] Christopher S. Carlson, Ryan O. Emerson, Anna M. Sherwood, Cindy Desmarais, Moon-Wook Chung, Joseph M. Parsons, Michelle S. Steen, Marissa A. LaMadrid-Herrmannsfeldt, David W. Williamson, Robert J. Livingston, David Wu, Brent L. Wood, Mark J. Rieder, and Harlan Robins. Using synthetic templates to design an unbiased multiplex PCR assay. *Nature Communications*, 4:2680, 2013.

[31] Jorge Carneiro, Kalet Leon, Íris Caramalho, Carline Van Den Dool, Rui Gardner, Vanessa Oliveira, Marie Louise Bergman, Nuno Sepúlveda, Tiago Paixão, Jose Faro, and Jocelyne Demengeot. When three is not a crowd: A Crossregulation Model of the dynamics and repertoire selection of regulatory CD4+ T cells. *Immunological Reviews*, 216(1):48–68, 2007.

[32] James A. Casbon, Robert J. Osborne, Sydney Brenner, and Conrad P. Lichtenstein. A method for counting PCR template molecules with application to next-generation sequencing. *Nucleic Acids Research*, 39(12):e81, 2011.

[33] Huabiao Chen, Zaza M Ndhlovu, Dongfang Liu, Lindsay C. Porter, Justin W. Fang, Sam Darko, Mark A. Brockman, Toshiyuki Miura, Zabrina L. Brumme, Arne Schneidewind, Alicja Piechocka-Trocha, Kevin T. Cesa, Jennifer Sela, Thai D. Cung, Ildiko Toth, Florencia Pereyra, Xu G. Yu, Daniel C. Douek, Daniel E. Kaufmann, Todd M. Allen, and Bruce D. Walker. TCR clonotypes modulate the protective effect of HLA class I molecules in HIV-1 infection. *Nature Immunology*, 13(7):691–700, 2012.

[34] Gary Cobbs. Stepwise kinetic equilibrium models of quantitative polymerase chain reaction. *BMC Bioinformatics*, 13(1):203, 2012.

[35] Rémi J Creusot, Lindy L Thomsen, John P Tite, and Benjamin M Chain. Local cooperation dominates over competition between CD4+ T cells of different antigen/MHC specificity. *Journal of Immunology*, 171:240–246, 2003.

[36] Pradyot Dash, Jennifer L. McClaren, Thomas H. Oguin, William Rothwell, Brandon Todd, Melissa Y. Morris, Jared Becksfort, Cory Reynolds, Scott A. Brown, Peter C. Doherty, and Paul G. Thomas. Paired analysis of TCRα and TCRβ chains at the single-cell level in mice. *Journal of Clinical Investigation*, 121(1):288–295, jan 2011.

[37] Mark M. Davis and Pamela J. Bjorkman. T-cell antigen receptor genes and T-cell recognition. *Nature*, 334(6181):395, 1988.

[38] Darren J. Day, Phyllis W. Speiser, Egbert Schulze, M. Bettendorf, Jodene Fitness, Francis Barany, and Perrin C. White. Identification of non-amplifying CYP21 genes when using PCR-based diagnosis of 21-hydroxylase deficiency in congenital adrenal hyperplasia (CAH) affected pedigrees. *Human Molecular Genetics*, 5(12):2039–2048, 1996.

[39] Rob J. De Boer and Alan S. Perelson. T Cell Repertoires and Competitive Exclusion. *Journal of Theoretical Biology*, 169:375–390, 1994.

[40] Stephane Demotz, Antonio Lanzavecchia, Eisel Ulrich, Heiner Niemann, Christian Widmann, and Giampietro Corradin. Delineation of several DR-restricted tetanus toxin T cell epitopes. *The Journal of Immunology*, 142(2):394–402, 1989.

[41] J R Dorfman, I Stefanová, K Yasutomo, and R N Germain. CD4+ T cell survival is not directly linked to self-MHC-induced TCR signaling. *Nature Immunology*, 1(4):329–335, 2000.

[42] Yuval Elhanati, Anand Murugan, Curtis G. Callan, Thierry Mora, and Aleksandra M. Walczak. Quantifying selection in immune receptor repertoires. *Proceedings of the National Academy of Sciences*, 111(27):9875–9880, 2014.

[43] J. S. Ellis, B. M. Chain, A. Cooke, M. A. Ibrahim, and D. R. Katz. Adjuvant Composition Determines the Induction of Type II Collagen-Induced Arthritis. *Scandinavian Journal of Immunology*, 36(1):49–56, 1992.

[44] Ryan O. Emerson, Anna M. Sherwood, Mark J. Rieder, Jamie Guenthoer, David W. Williamson, Christopher S. Carlson, Charles W. Drescher, Muneesh Tewari, Jason H. Bielas, and Harlan S. Robins. High-throughput sequencing of T-cell receptors reveals a homogeneous repertoire of tumour-infiltrating lymphocytes in ovarian cancer. *The Journal of Pathology*, 231(4):433–440, 2013.

[45] J Douglas Freeman, René L Warren, John R Webb, Brad H Nelson, and Robert a Holt. Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. *Genome research*, 19(10):1817–1824, 2009.

[46] Olivier Gaide, Ryan O Emerson, Xiaodong Jiang, Nicholas Gulati, Suzanne Nizza, Cindy Desmarais, Harlan Robins, James G Krueger, Rachael A Clark, and Thomas S Kupper. Common clonal origin of central and resident memory T cells following skin immunization. *Nature Medicine*, 21(6):647–653, 2015.

[47] Nicholas RJ Gascoigne and Oreste Acuto. THEMIS: a critical TCR signal regulator for ligand discrimination. *Current Opinion in Immunology*, 33:86–92, 2015.

[48] Martin Gellert. V(D)J Recombination: RAG Proteins, Repair Factors and Regulation. *Annual Review of Biochemistry*, 71(1):101–132, 2002.

[49] Jana L Gevertz, Stanley M Dunn, and Charles M Roth. Mathematical model of real-time PCR kinetics. *Biotechnology and Bioengineering*, 92(3):346–355, 2005.

[50] Véronique Giudicelli, Denys Chaume, and Marie-Paule Lefranc. IMGT/GENE-DB: a comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Research*, 33:D256–61, 2005.

[51] Jack Gorski, Teresa Piatek, Maryam Yassai, Jessica Gorski, and Krystyna Maslanka. Improvements in Repertoire Analysis by CDR3 Size Spectratyping. *Annals of the New York Academy of Sciences*, 756:99–102, 1995.

[52] Jack Gorski, Maryam Yassai, Xiaolei Zhu, Bret Kisella, Carolyn Keever, and Neal Fiomenberg. Circulating T cell repertoire complexity in normal individuals and bone marrow recipients analyzed by CDR3 size spectratyping. Correlation with immune status. *Journal of Immunology*, 152(10):5109 – 5119, 1994.

[53] Alena Gros, Paul F Robbins, Xin Yao, Yong F Li, Simon Turcotte, Eric Tran, John R Wunderlich, Arnold Mixon, Shawn Farid, Mark E Dudley, Ken-ichi Hanada, Jorge R Almeida, Sam Darko,

Daniel C Douek, James C Yang, and Steven A Rosenberg. PD-1 identifies the patient-specific CD8+ tumor-reactive repertoire infiltrating human tumors. *Journal of Clinical Investigation*, 124(5):2246–2259, 2014.

[54] Bradley Hall, John M Micheletti, Pooja Satya, Krystal Ogle, Jack Pollard, and Andrew D Ellington. Design, Synthesis, and Amplification of DNA Pools for In Vitro Selection. In *Current Protocols in Molecular Biology*, pages 24.2.1–24.2.27. John Wiley & Sons, Inc., Hoboken, NJ, USA, 2009.

[55] Arnold Han, Jacob Glanville, Leo Hansmann, and Mark M Davis. Linking T-cell receptor sequence to functional phenotype at the single-cell level. *Nature Biotechnology*, 32(7):684–692, 2014.

[56] Bret Hanlon and Anand N. Vidyashankar. Inference for Quantitation Parameters in Polymerase Chain Reactions via Branching Processes With Random Effects. *Journal of the American Statistical Association*, 106(494):525–533, 2011.

[57] Jason Hataye, James J. Moon, Alexander Khoruts, Cavan Reilly, and Marc K. Jenkins. Naive and Memory CD4+ T Cell Survival Controlled by Clonal Abundance. *Science*, 312(5770):114–116, 2006.

[58] D Hawiger, K Inaba, Y Dorsett, M Guo, K Mahnke, M Rivera, J V Ravetch, R M Steinman, and M C Nussenzweig. Dendritic cells induce peripheral T cell unresponsiveness under steady state conditions in vivo. *The Journal of Experimental Medicine*, 194(6):769–779, 2001.

[59] James M. Heather, Katharine Best, Theres Oakes, Eleanor R. Gray, Jennifer K. Roe, Niclas Thomas, Nir Friedman, Mahdad Noursadeghi, and Benjamin Chain. Dynamic Perturbations of the T-Cell Receptor Repertoire in Chronic HIV Infection and following Antiretroviral Therapy. *Frontiers in Immunology*, 6:644, 2016.

[60] Arne Henningsen and Ott Toomet. maxLik: A package for maximum likelihood estimation in R. *Computational Statistics*, 26(3):443–458, 2011.

[61] Kristin A. Hogquist, Troy A. Baldwin, and Stephen C. Jameson. Central tolerance: learning self-control in the thymus. *Nature Reviews Immunology*, 5(10):772–782, 2005.

[62] Christopher J. Holland, David K. Cole, and Andrew Godkin. Re-directing CD4+ T cell responses with the flanking residues of MHC class II-bound peptides: the core is not enough. *Frontiers in Immunology*, 4:172, 2013.

[63] B. Howie, A. M. Sherwood, A. D. Berkebile, J. Berka, R. O. Emerson, D. W. Williamson, I. Kirsch, M. Vignali, M. J. Rieder, C. S. Carlson, and H. S. Robins. High-throughput pairing of T cell receptor alpha and beta sequences. *Science Translational Medicine*, 7:301ra131, 2015.

[64] H Hsu, J Xiong, and D V Goeddel. The TNF receptor 1-associated protein TRADD signals cell death and NF-kappa B activation. *Cell*, 81:495–504, 1995.

[65] Mohammad A.A Ibrahim, Benjamin M. Chain, and David R. Katz. The injured cell: The role of the dendritic cell system as a sentinel receptor pathway. *Immunology Today*, 16:181–186, 1995.

[66] Botond Z Igyártó and Daniel H Kaplan. Antigen presentation by Langerhans cells. *Current Opinion in Immunology*, 25:115–119, 2013.

[67] Takashi Izawa, Tomoyuki Kondo, Mie Kurosawa, Ritsuko Oura, Kazuma Matsumoto, Eiji Tanaka, Akiko Yamada, Rieko Arakaki, Yasusei Kudo, Yoshio Hayashi, and Naozumi Ishimaru. Fas-Independent T-Cell Apoptosis by Dendritic Cells Controls Autoimmune Arthritis in MRL/lpr Mice. *PLoS ONE*, 7(12), 2012.

[68] C Jacob and J Peccoud. Estimation of the parameters of a branching process from migrating binomial observations. *Advances in Applied Probability*, 30(4):948–967, 1998.

[69] Charles A Janeway. Approaching the asymptote? Evolution and revolution in immunology. *Coldspring Harbor Symposia on Quantitative Biology*, 54(9):1–13, 1989.

[70] Charles A. Janeway and Ruslan Medzhitov. Innate immune recognition. *Annual Review of Immunology*, 20(2):197–216, 2002.

[71] Jeffrey L. Jorgensen, Ursula Esser, Barbara Fazekas de St Groth, Philip A. Reay, and Mark M. Davis. Mapping T-cell receptor-peptide contacts by variant peptide immunization of single-chain transgenics. *Nature*, 355(6357):224–230, 1992.

[72] J W Kappler, N Roehm, and P Marrack. T cell tolerance by clonal elimination in the thymus. *Cell*, 49(2):273–280, 1987.

[73] Yong Ke, Ying Li, and Judith A Kapp. Ovalbumin injected with complete Freund's adjuvant stimulates cytolytic responses. *European Journal of Immunology*, 25(2):549–53, 1995.

[74] S. Khailaie, P. A. Robert, A. Toker, J. Huehn, and M. Meyer-Hermann. A Signal Integration Model of Thymic Selection and Natural Regulatory T Cell Commitment. *The Journal of Immunology*, 193:5983–5996, 2014.

[75] Marek Kimmel and David E Axelrod. *Branching Processes in Biology*. Springer-Verlag New York, 2002.

[76] I. R. Kirsch, R. Watanabe, J. T. O'Malley, D. W. Williamson, L.-L. Scott, C. P. Elco, J. E. Teague, A. Gehad, E. L. Lowry, N. R. LeBoeuf, J. G. Krueger, H. S. Robins, T. S. Kupper, and R. A. Clark. TCR sequencing facilitates diagnosis and identifies mature T cells as the cell of origin in CTCL. *Science Translational Medicine*, 7:308ra158, 2015.

[77] Teemu Kivioja, Anna Vähärautio, Kasper Karlsson, Martin Bonke, Martin Enge, Sten Linnarsson, and Jussi Taipale. Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods*, 9(1):72–74, 2011.

[78] P. L. Klarenbeek, M. J. H. de Hair, M. E. Doorenspleet, B. D. C. van Schaik, R. E. E. Esveldt, M. G. H. van de Sande, T. Cantaert, D. M. Gerlag, D. Baeten, A. H. C. van Kampen, F. Baas, P. P. Tak, and N. de Vries. Inflamed target tissue provides a specific niche for highly expanded T-cell clones in early human autoimmune disease. *Annals of the Rheumatic Diseases*, 71(6):1088–1093, 2012.

[79] Paul L Klarenbeek, Paul P Tak, Barbera D C van Schaik, Aeilko H Zwinderman, Marja E Jakobs, Zhuoli Zhang, Antoine H C van Kampen, René A W van Lier, Frank Baas, and Niek de Vries. Human T-cell memory consists mainly of unexpanded clones. *Immunology Letters*, 133(1):42–48, 2010.

[80] Ludger Klein, Maria Hinterberger, Gerald Wirnsberger, and Bruno Kyewski. Antigen presentation in the thymus for positive selection and central tolerance induction. *Nature reviews. Immunology*, 9(12):833–44, 2009.

[81] Mark Klinger, Katherine Kong, Martin Moorhead, Li Weng, Jianbiao Zheng, and Malek Faham. Combining next-generation sequencing and immune assays: a novel method for identification of antigen-specific T cells. *PloS ONE*, 8(9):e74231, 2013.

[82] Michelle Krogsgaard and Mark M Davis. How T cells 'see' antigen. *Nature Immunology*, 6(3):239–245, 2005.

[83] Shinya Kurata, Takahiro Kanagawa, Y. Magariyama, K. Takatsu, K. Yamada, T. Yokomaku, and Y. Kamagata. Reevaluation and Reduction of a PCR Bias Caused by Reannealing of Templates. *Applied and Environmental Microbiology*, 70(12):7545–7549, 2004.

[84] N. Lalam, C. Jacob, and P. Jagers. Modelling the PCR Amplification Process by a Size-Dependent Branching Process and Estimation of the Efficiency. *Advances in Applied Probability*, 36(2):602–615, 2004.

[85] H. Li, C. Ye, G. Ji, X. Wu, Z. Xiang, Y. Li, Y. Cao, X. Liu, D. C. Douek, D. A. Price, and J. Han. Recombinatorial Biases and Convergent Recombination Determine Interindividual TCR Sharing in Murine Thymocytes. *The Journal of Immunology*, 189(5):2404–2413, 2012.

[86] Shuo Li, Marie-Paule Lefranc, John J Miles, Eltaf Alamyar, Véronique Giudicelli, Patrice Duroux, J Douglas Freeman, Vincent D A Corbin, Jean-Pierre Scheerlinck, Michael A Frohman, Paul U Cameron, Magdalena Plebanski, Bruce Loveland, Scott R Burrows, Anthony T Papenfuss, and Eric J Gowans. IMGT/HighV QUEST paradigm for T cell receptor IMGT clonotype diversity and next generation repertoire immunoprofiling. *Nature Communications*, 4:2333, 2013.

[87] Aaron C. Logan, Nikita Vashi, Malek Faham, Victoria Carlton, Katherine Kong, Ismael Buño, Jianbiao Zheng, Martin Moorhead, Mark Klinger, Bing Zhang, Amna Waqar, James L. Zehnder, and David B. Miklos. Immunoglobulin and T cell receptor gene high-throughput sequencing quantifies minimal residual disease in acute lymphoblastic leukemia and predicts post-transplantation relapse and survival. *Biology of Blood and Marrow Transplantation*, 20(9):1307–1313, 2014.

[88] Wei Luo, Jin Su, Xiao-Bing Zhang, Zhi Yang, Ming-Qian Zhou, Zhen-Min Jiang, Pei-Pei Hao, Su-Dong Liu, Qian Wen, Qi Jin, and Li Ma. Limited T Cell Receptor Repertoire Diversity in Tuberculosis Patients Correlates with Clinical Severity. *PLoS ONE*, 7(10):e48117, 2012.

[89] Manfred B Lutz, Anja Döhler, and Hiroaki Azukizawa. Revisiting the tolerogenicity of epidermal Langerhans cells. *Immunology and Cell Biology*, 88(4):381–386, 2010.

[90] Grant Lythe, Robin E. Callard, Rollo Hoare, and Carmen Molina-París. How many TCR clonotypes does a body maintain? *Journal of Theoretical Biology*, 389:214–224, 2015.

[91] Asaf Madi, Eric Shifrut, Shlomit Reich-Zeliger, Hilah Gal, Katharine Best, Wilfred Ndifon, Benjamin Chain, Irun R Cohen, and Nir Friedman. T-cell receptor repertoires share a restricted set of public and abundant CDR3 sequences that are associated with self-related immunity. *Genome Research*, 24(10):1603–1612, 2014.

[92] Roberto A. Maldonado and Ulrich H. von Andrian. How Tolerogenic Dendritic Cells Induce Regulatory T Cells. In *Advances in Immunology*, volume 108, pages 111–165. Elsevier Inc, 2010.

[93] Ilgar Z Mamedov, Olga V Britanova, Ivan V Zvyagin, Maria A Turchaninova, Dmitriy A Bolotin, Ekaterina V Putintseva, Yuriy B Lebedev, and Dmitriy M Chudakov. Preparing Unbiased T-Cell Receptor and Antibody cDNA Libraries for the Deep Next Generation Sequencing Profiling. *Frontiers in Immunology*, 4:456, 2013.

[94] Marcel Margulies, Michael Egholm, William E Altman, Said Attiya, Joel S Bader, Lisa A Bemben, Jan Berka, Michael S Braverman, Yi-Ju Chen, Zhoutao Chen, Scott B Dewell, Lei Du, Joseph M Fierro, Xavier V Gomes, Brian C Godwin, Wen He, Scott Helgesen, Chun He Ho, Gerard P Irzyk, Szilveszter C Jando, Maria L I Alenquer, Thomas P Jarvie, Kshama B Jirage, Jong-Bum Kim, James R Knight, Janna R Lanza, John H Leamon, Steven M Lefkowitz, Ming Lei, Jing Li, Kenton L Lohman, Hong Lu, Vinod B Makhijani, Keith E McDade, Michael P McKenna, Eugene W Myers, Elizabeth Nickerson, John R Nobile, Ramona Plant, Bernard P Puc, Michael T Ronan, George T Roth, Gary J Sarkis, Jan Fredrik Simons, John W Simpson, Maithreyan Srinivasan, Karrie R Tartaro, Alexander Tomasz, Kari A Vogt, Greg A Volkmer, Shally H. Wang, Yong Wang, Michael P Weiner, Pengguang Yu, Richard F Begley, and Jonathan M Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–80, 2005.

[95] Polly Matzinger. Tolerance, Danger, and the Extended Family. *Annual Review of Immunology*, 12(1):991–1045, 1994.

[96] Polly Matzinger. The danger model: a renewed sense of self. *Science*, 296:301–305, 2002.

[97] Andreas Mayer, Vijay Balasubramanian, Thierry Mora, and Aleksandra M Walczak. How a well-adapted immune system is organized. *Proceedings of the National Academy of Sciences*, 112(19):5950–5955, 2015.

[98] Megan L McCloskey, Reinhard Stöger, R Scott Hansen, and Charles D Laird. Encoding PCR Products with Batch-stamps and Barcodes. *Biochemical Genetics*, 45(11-12):761–767, 2007.

[99] Jonathan R McDaniel, Brandon J DeKosky, Hidetaka Tanno, Andrew D Ellington, and George Georgiou. Ultra-high-throughput sequencing of the immune receptor repertoire from millions of lymphocytes. *Nature Protocols*, 11(3):429–442, 2016.

[100] Ilhem Messaoudi, Jose A. Guevara Patino, Ruben Dyall, Joel LeMaoult, and Janko Nikolich-Zugich. Direct Link Between mhc Polymorphism, T Cell Avidity, and Diversity in Immune Defense. *Science*, 298(5599):1797–1800, 2002.

[101] Everett H. Meyer, Andro R. Hsu, Joanna Liliental, A. Lohr, Mareike Florek, James L. Zehnder, Sam Strober, Philip Lavori, David B. Miklos, David S. Johnson, and Robert S. Negrin. A distinct evolution of the T-cell repertoire categorizes treatment refractory gastrointestinal acute graft-versus-host disease. *Blood*, 121(24):4955–4962, 2013.

[102] John J Miles, Daniel C Douek, and David a Price. Bias in the $\alpha\beta$ T-cell repertoire: implications for disease pathogenesis and vaccination. *Immunology and Cell Biology*, 89(3):375–87, 2011.

[103] Mark J Miller, Arsalan S Hejazi, Sindy H Wei, Michael D Cahalan, and Ian Parker. T cell repertoire scanning is promoted by dynamic dendritic cell behavior and random T cell motility in the lymph node. *Proceedings of the National Academy of Sciences*, 101(4):998–1003, 2004.

[104] Brooks E Miner, Reinhard J Stoger, Alice F Burden, Charles D Laird, and R Scott Hansen. Molecular barcodes detect redundancy and contamination in hairpin-bisulfite PCR. *Nucleic Acids Research*, 32(17):e135, 2004.

[105] James R. Moore. The benefits of diversity: Heterogenous DC populations allow for both immunity and tolerance. *Journal of Theoretical Biology*, 357:86–102, 2014.

[106] Gerald P Morris and Paul M Allen. How the TCR balances sensitivity and specificity for the recognition of self and pathogens. *Nature immunology*, 13(2):121–8, 2012.

[107] Kenneth Murphy, Paul Travers, and Mark Walport. *Janeway's Immunobiology*. Garland Science, 7th edition, 2008.

[108] Anand Murugan, Thierry Mora, Aleksandra M Walczak, and Curtis G Callan. Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proceedings of the National Academy of Sciences*, 109(40):16161–6, 2012.

[109] N. Naderi, S. M. Moazzeni, A. A. Pourfathollah, and K. Alimoghaddam. High expression of fas ligand on cord blood dendritic cells: A possible immunoregulatory mechanism after cord blood transplantation. *Transplantation Proceedings*, 43(10):3913–3919, 2011.

[110] Ugrappa Nagalakshmi, Zhong Wang, Karl Waern, Chong Shou, Debasish Raha, Mark Gerstein, and Michael Snyder. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science*, 320(5881):1344–1349, 2008.

[111] W. Ndifon, H. Gal, E. Shifrut, R. Aharoni, N. Yissachar, N. Waysbort, S. Reich-Zeliger, R. Arnon, and N. Friedman. Chromatin conformation governs T-cell receptor J gene segment usage. *Proceedings of the National Academy of Sciences*, 109:15865–15870, 2012.

[112] D Nesić and S Vukmanović. MHC class I is required for peripheral accumulation of CD8+ thymic emigrants. *Journal of Immunology*, 160(8):3705–12, 1998.

[113] Phuong Nguyen, Jing Ma, Deqing Pei, Caroline Obert, Cheng Cheng, and Terrence L Geiger. Identification of errors introduced during high throughput sequencing of the T cell receptor repertoire. *BMC Genomics*, 12(1):106, 2011.

[114] Florence Paillard, Ghislaine Sterkers, Georges Bismutha, Elisabeth Gomard, and Catherine Vaquero. Lymphokine mRNA and T cell multireceptor mRNA of the Ig super gene family are reciprocally modulated during human T cell activation. *European Journal of Immunology*, 18(18):1643–1646, 1988.

[115] Ed Palmer and Dieter Naeher. Affinity threshold for thymic selection through a T-cell receptor-co-receptor zipper. *Nature reviews Immunology*, 9(3):207–13, 2009.

[116] Wenjing Pan, Miranda Byrne-Steele, Chunlin Wang, Stanley Lu, Scott Clemmons, Robert J Zahorchak, and Jian Han. DNA polymerase preference determines PCR priming efficiency. *BMC Biotechnology*, 14(1):10, 2014.

[117] J Peccoud and C Jacob. Theoretical uncertainty of measurements using quantitative polymerase chain reaction. *Biophysical Journal*, 71(1):101–108, 1996.

[118] E Pienaar, M Theron, M Nelson, and HJ Viljoen. A quantitative model of error accumulation during PCR amplification. *Computational Biology and Chemistry*, 30(2):102–111, 2006.

[119] Martin F. Polz and Colleen M. Cavanaugh. Bias in template-to-product ratios in multitemplate PCR. *Applied and Environmental Microbiology*, 64(10):3724–3730, 1998.

[120] Qian Qi, Yi Liu, Yong Cheng, Jacob Glanville, David Zhang, Ji-Yeun Lee, Richard a Olshen, Cornelia M Weyand, Scott D Boyd, and Jörg J Goronzy. Diversity and clonal selection in the human T-cell repertoire. *Proceedings of the National Academy of Sciences*, 2014.

[121] J P Ridge, F Di Rosa, and P Matzinger. A conditioned dendritic cell can be a temporal bridge between a CD4+ T-helper and a T-killer cell. *Nature*, 393(6684):474–478, 1998.

[122] Lidia Robert, Jennifer Tsoi, Xiaoyan Wang, Ryan Emerson, Blanca Homet, Thinle Chodon, Stephen Mok, Rong Rong Huang, Alistair J. Cochran, Begoña Comin-Anduix, Richard C. Koya, Thomas G. Graeber, Harlan Robins, and Antoni Ribas. CTLA4 Blockade Broadens the Peripheral T-Cell Receptor Repertoire. *Clinical Cancer Research*, 20(9):2424–2432, 2014.

[123] J M Robertson, P E Jensen, and B D Evavold. DO11.10 and OT-II T cells recognize a C-terminal ovalbumin 323-339 epitope. *Journal of Immunology*, 164(9):4706–4712, 2000.

[124] Harlan Robins, Cindy Desmarais, Jessica Matthis, Robert Livingston, Jessica Andriesen, Helena Reijonen, Christopher Carlson, Gerold Nepom, Cassian Yee, and Karen Cerosaletti. Ultra-sensitive detection of rare T cell clones. *Journal of Immunological Methods*, 375(1-2):14–19, 2012.

[125] Harlan S Robins, Paulo V Campregher, Santosh K Srivastava, Abigail Wacher, Cameron J Turtle, Orsalem Kahsai, Stanley R Riddell, Edus H Warren, and Christopher S Carlson. Comprehensive assessment of T-cell receptor β-chain diversity in $\alpha\beta$ T cells. *Blood*, 114(19):4099–4107, 2009.

[126] Harlan S Robins, Santosh K Srivastava, Paulo V Campregher, Cameron J Turtle, Jessica Andriesen, Stanley R Riddell, Christopher S Carlson, and Edus H Warren. Overlap and Effective Size of the Human CD8+ T Cell Receptor Repertoire. *Science Translational Medicine*, 2:47ra64, 2010.

[127] Eitan Rubin and Avraham A. Levy. A mathematical model and a computerized simulation of PCR using complex templates. *Nucleic Acids Research*, 24(18):3538–3545, 1996.

[128] Shimon Sakaguchi. Naturally Arising CD4+ Regulatory T Cells for Immunologic Self-Tolerance and Negative Control of Immune Responses. *Annual Review of Immunology*, 22(1):531–562, 2004.

[129] Shimon Sakaguchi, Tomoyuki Yamaguchi, Takashi Nomura, and Masahiro Ono. Regulatory T Cells and Immune Tolerance. *Cell*, 133(5):775–787, 2008.

[130] Claude Edmund Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[131] Anna M Sherwood, Ryan O Emerson, Dominique Scherer, Nina Habermann, Katharina Buck, Jürgen Staffa, Cindy Desmarais, Niels Halama, Dirk Jaeger, Peter Schirmacher, Esther Herpel, Matthias Kloor, Alexis Ulrich, Martin Schneider, Cornelia M Ulrich, and Harlan Robins. Tumor-infiltrating lymphocytes in colorectal tumors display a diversity of T cell receptor sequences that differ from the T cells in adjacent mucosal tissue. *Cancer immunology immunotherapy*, 2013.

[132] K. Shiroguchi, T. Z. Jia, P. A. Sims, and X. S. Xie. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proceedings of the National Academy of Sciences*, 109(4):1347–1352, 2012.

[133] Mikhail Shugay, Olga V Britanova, Ekaterina M Merzlyak, Maria a Turchaninova, Ilgar Z Mamedov, Timur R Tuganbaev, Dmitriy A Bolotin, Dmitry B Staroverov, Ekaterina V Putintseva, Karla Plevova, Carsten Linnemann, Dmitriy Shagin, Sarka Pospisilova, Sergey Lukyanov, Ton N Schumacher, and Dmitriy M Chudakov. Towards error-free profiling of immune repertoires. *Nature Methods*, 11(6):653–5, 2014.

[134] Ralph M Steinman. The Dendritic Cell System and its Role in Immunogenicity. *Annual Review of Immunology*, 9:271–96, 1991.

[135] Ralph M. Steinman, Daniel Hawiger, and Michel C. Nussenzweig. Tolerogenic Dendritic Cells. *Annual Review of Immunology*, 21(1):685–711, 2003.

[136] Ondrej Stepanek, Arvind S Prabhakar, Celine Osswald, Carolyn G King, Anna Bulek, Dieter Naeher, Marina Beaufils-hugot, Michael L Abanto, Virginie Galati, Barbara Hausmann, Rosemarie Lang, David K Cole, Eric S Huseby, Andrew K Sewell, Arup K Chakraborty, and Ed Palmer. Coreceptor Scanning by the T Cell Receptor Provides a Mechanism for T Cell Tolerance. *Cell*, pages 1–13, 2014.

[137] Emily R Stirk, Grant Lythe, Hugo A van den Berg, and Carmen Molina-París. Stochastic competitive exclusion in the maintenance of the naïve T cell repertoire. *Journal of Theoretical Biology*, 265(3):396–410, 2010.

[138] Emily R Stirk, Carmen Molina-París, and Hugo A van den Berg. Stochastic niche structure and diversity maintenance in the T cell repertoire. *Journal of Theoretical Biology*, 255(2):237–49, 2008.

[139] G Stolovitzky and G Cecchi. Efficiency of DNA replication in the polymerase chain reaction. *Proceedings of the National Academy of Sciences*, 93(23):12947–12952, 1996.

[140] Kari E. Sufficool, Christina M. Lockwood, Haley J. Abel, Ian S. Hagemann, Jonathan A. Schumacher, Todd W. Kelley, and Eric J. Duncavage. T-cell clonality assessment by next-generation sequencing improves detection sensitivity in mycosis fungoides. *Journal of the American Academy of Dermatology*, 73(2):228–236.e2, 2015.

[141] Kensuke Takada and Stephen C Jameson. Naive T cell homeostasis: from awareness of space to a sense of place. *Nature reviews Immunology*, 9(12):823–32, 2009.

[142] S Takeda, H R Rodewald, H Arakawa, H Bluethmann, and T Shimizu. MHC class II molecules are not required for survival of newly generated CD4+ T cells, but affect their long-term life span. *Immunity*, 5(3):217–28, 1996.

[143] J T Tan, E Dudl, E LeRoy, R Murray, J Sprent, K I Weinberg, and C D Surh. IL-7 is critical for homeostatic proliferation and survival of naive T cells. *Proceedings of the National Academy of Sciences*, 98(15):8732–7, 2001.

[144] Niclas Thomas, Katharine Best, Mattia Cinelli, Shlomit Reich-Zeliger, Hilah Gal, Eric Shifrut, Asaf Madi, Nir Friedman, John Shawe-Taylor, and Benny Chain. Tracking global changes induced in the CD4 T-cell receptor repertoire by immunization with a complex antigen using short stretches of CDR3 protein sequence. *Bioinformatics*, 30(22):3181–3188, 2014.

[145] Niclas Thomas, James M Heather, Wilfred Ndifon, John Shawe-Taylor, and Benjamin Chain. Decombinator: a tool for fast, efficient gene assignment in T-cell receptor sequences using a finite state machine. *Bioinformatics*, 29(5):542–550, 2013.

[146] Niclas Thomas, Lenka Matejovicova, Wichat Srikusalanukul, John Shawe-Taylor, and Benny Chain. Directional Migration of Recirculating Lymphocytes through Lymph Nodes via Random Walks. *PLoS ONE*, 7(9), 2012.

[147] Paul C. Tumeh, Christina L. Harview, Jennifer H. Yearley, I. Peter Shintaku, Emma J. M. Taylor, Lidia Robert, Bartosz Chmielowski, Marko Spasic, Gina Henry, Voicu Ciobanu, Alisha N. West, Manuel Carmona, Christine Kivork, Elizabeth Seja, Grace Cherry, Antonio J. Gutierrez, Tristan R. Grogan, Christine Mateus, Gorana Tomasic, John A. Glaspy, Ryan O. Emerson, Harlan Robins, Robert H. Pierce, David A. Elashoff, Caroline Robert, and Antoni Ribas. PD-1 blockade induces responses by inhibiting adaptive immune resistance. *Nature*, 515(7528):568–571, 2014.

[148] Maria A Turchaninova, Olga V Britanova, Dmitriy A Bolotin, Mikhail Shugay, Ekaterina V Putintseva, Dmitriy B Staroverov, George Sharonov, Dmitriy Shcherbo, Ivan V Zvyagin, Ilgar Z Mamedov, Carsten Linnemann, Ton N Schumacher, and Dmitriy M Chudakov. Pairing of T-cell receptor chains via emulsion PCR. *European Journal of Immunology*, 43(9):2507–15, 2013.

[149] Shannon J Turley, Anne L Fletcher, and Kutlu G Elpek. The stromal and haematopoietic antigen-presenting cells that reside in secondary lymphoid organs. *Nature Reviews Immunology*, 10(12):813–825, 2010.

[150] Emil R Unanue. Antigen-presenting function of the macrophage. *Annual review of immunology*, 2:395–428, 1984.

[151] Jeroen W J van Heijst, Izaskun Ceberio, Lauren B Lipuma, Dane W Samilo, Gloria D Wasilewski, Anne Marie R Gonzales, Jimmy L Nieves, Marcel R M van den Brink, Miguel A Perales, and Eric G Pamer. Quantitative assessment of T cell repertoire recovery after hematopoietic stem cell transplantation. *Nature Medicine*, 19(3):372–7, 2013.

[152] Vanessa Venturi, Katherine Kedzierska, David A Price, Peter C Doherty, Daniel C Douek, Stephen J Turner, and Miles P Davenport. Sharing of T cell receptors in antigen-specific responses is driven by convergent recombination. *Proceedings of the National Academy of Sciences*, 103(49):18691–18696, 2006.

[153] Vanessa Venturi, Máire F Quigley, Hui Yee Greenaway, Pauline C Ng, Zachary S Ende, Tina McIntosh, Tedi E Asher, Jorge R Almeida, Samuel Levy, David A Price, Miles P Davenport, and Daniel C Douek. A mechanism for TCR sharing between T cell subsets and individuals revealed by pyrosequencing. *Journal of Immunology*, 186(7):4285–4294, 2011.

[154] Christopher Vollmers, Rene V Sit, Joshua A Weinstein, Cornelia L Dekker, and Stephen R Quake. Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proceedings of the National Academy of Sciences*, 110(33):13463–13468, 2013.

[155] C. Wang, Catherine M Sanders, Qunying Yang, Harry W Schroeder, Elijah Wang, Farbod Babrzadeh, Baback Gharizadeh, Richard M Myers, James R Hudson, Ronald W Davis, and Jian Han. High throughput sequencing reveals a complex pattern of dynamic interrelationships among human T cell subsets. *Proceedings of the National Academy of Sciences*, 107(4):1518–1523, 2010.

[156] René L Warren, J Douglas Freeman, Thomas Zeng, Gina Choe, Sarah Munro, Richard Moore, John R Webb, and Robert A Holt. Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. *Genome Research*, 21(5):790–797, 2011.

[157] Joshua A. Weinstein, Ning Jiang, Richard A. White, Daniel S Fisher, and Stephen R Quake. High-Throughput Sequencing of the Zebrafish Antibody Repertoire. *Science*, 324(5928):807–810, 2009.

[158] Scott E Whitney, Alugupally Sudhir, R.Michael Nelson, and Hendrik J Viljoen. Principles of rapid polymerase chain reactions: mathematical modeling and experimental verification. *Computational Biology and Chemistry*, 28(3):195–209, 2004.

[159] Linda Wooldridge, Julia Ekeruche-Makinde, Hugo A. Van Den Berg, Anna Skowera, John J. Miles, Mai Ping Tan, Garry Dolton, Mathew Clement, Sian Llewellyn-Lacey, David A. Price, Mark Peakman, and Andrew K. Sewell. A single autoimmune T cell receptor recognizes more than a million different peptides. *Journal of Biological Chemistry*, 287(2):1168–1177, 2012.

[160] D Wu, A Sherwood, J R Fromm, S S Winter, K P Dunsmore, M L Loh, H A Greisman, D E Sabath, B L Wood, and H Robins. High-Throughput Sequencing Detects Minimal Residual Disease in Acute T Lymphoblastic Leukemia. *Science Translational Medicine*, 4:134ra63, 2012.

[161] Qiong Wu, Anne M Pesenacker, Alka Stansfield, Douglas King, Dawn Barge, Helen E Foster, Mario Abinun, and Lucy R Wedderburn. Immunological characteristics and T-cell receptor clonal diversity in children with systemic juvenile idiopathic arthritis undergoing T-cell-depleted autologous stem cell transplantation. *Immunology*, 142(2):227–236, 2014.

[162] Xiongfei Xu, Hai Yi, Zhenhong Guo, Cheng Qian, Sheng Xia, Yushi Yao, and Xuetao Cao. Splenic Stroma-Educated Regulatory Dendritic Cells Induce Apoptosis of Activated CD4 T Cells via FasL-Enhanced IFN-γ and Nitric Oxide. *The Journal of Immunology*, 2011.

[163] Wong Yu, Ning Jiang, PeterJ.R. Ebert, BrianA. Kidd, Sabina Müller, PederJ. Lund, Jeremy Juang, Keishi Adachi, Tiffany Tse, MichaelE. Birnbaum, EvanW. Newell, DarrellM. Wilson, GijsbertM. Grotenbreg, Salvatore Valitutti, StephenR. Quake, and MarkM. Davis. Clonal Deletion Prunes but Does Not Eliminate Self-Specific $\alpha\beta$ CD8+ T Lymphocytes. *Immunity*, 42(5):929–941, 2015.

[164] Xiaomin Yu, Jorge R Almeida, Sam Darko, Mirjam van der Burg, Suk See DeRavin, Harry Malech, Andrew Gennery, Ivan Chinn, Mary Louise Markert, Daniel C Douek, and Joshua D Milner. Human syndromes of immunodeficiency and dysregulation are characterized by distinct defects in T-cell receptor repertoire development. *The Journal of Allergy and Clinical Immunology*, 133(4):1109–15, 2014.

[165] Xu G Yu, Mathias Lichterfeld, Katie L Williams, Javier Martinez-Picado, and Bruce D Walker. Random T-cell receptor recruitment in human immunodeficiency virus type 1 (HIV-1)-specific CD8+ T cells from genetically identical twins infected with the same HIV-1 strain. *Journal of Virology*, 81(22):12666–9, 2007.

[166] Lior Zangi, Yael Zlotnikov Klionsky, Liran Yarimi, Esther Bachar-Lustig, Yaki Eidelstein, Elias Shezen, David Hagin, Yumi Ito, Toshiyuki Takai, Shlomit Reich-Zeliger, Assaf Lask, Oren Milstein, Steffen Jung, Vera Shinder, and Yair Reisner. Deletion of cognate CD8 T cells by immature dendritic cells: A novel role for perforin, granzyme A, TREM-1, and TLR7. *Blood*, 120(8):1647–1657, 2012.

[167] Yael Zlotnikov-Klionsky, Bar Nathansohn-Levi, Elias Shezen, Chava Rosen, Sivan Kagan, Liat Bar-On, Steffen Jung, Eric Shifrut, Shlomit Reich-Zeliger, Nir Friedman, Rina Aharoni, Ruth Arnon, Oren Yifa, Anna Aronovich, and Yair Reisner. Perforin-Positive Dendritic Cells Exhibit an Immuno-regulatory Role in Metabolic Syndrome and Autoimmunity. *Immunity*, 43(4):776–787, 2015.

Appendices

Appendix A

PCR amplification heterogeneity: Supplementary Information

A.1 Empirical distribution of barcodes

Aggregating barcode data from a number of sequence runs allows us to infer information about the structure of the pool of barcode-oligos. We counted barcode labelling events across our experiments (Figure A.3a, black dots) and found that the majority of barcodes we observed have only been seen in one labelling event while some have been observed up to 12 times. If the pool of barcodes was uniformly distributed we would expect the barcode labelling event counts to be zero-truncated Poisson distribution, while if the pool follows some other distribution we expect the barcode labelling event counts to follow a zero-truncated mixed Poisson distribution. We fitted, via maximum likelihood, zero-truncated Poisson and mixed Poisson distributions, and saw that the zero-truncated Poisson mixed with a lognormal distribution provided the best fit to the observed barcode labelling event counts (Figure A.3a, coloured lines). This suggests that the pool of barcodes we are labelling our molecules from is lognormally distributed. Zero-truncated mixed Poisson models are fitted to data via maximum likelihood optimisation in R. Maximum likelihood optimisation was performed using the maxNR of the MaxLik package [60].

From the fitted parameters of the zero-truncated Poisson-lognormal distribution we are

able to infer the structure of the available barcode pool (Figure A.3b), showing that in a pool of 10^8 barcode-oligos most labels occur fewer than 5 times while a small proportion occur up to 30 times.

A.2 Supplementary Figures

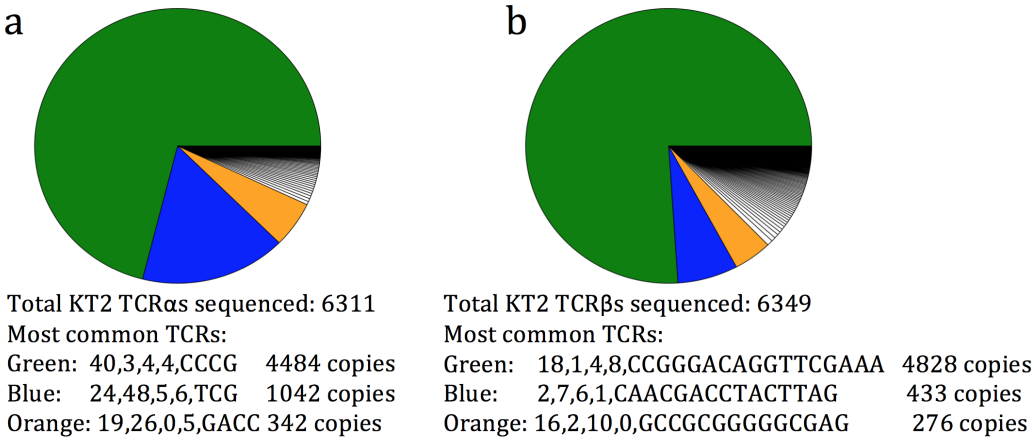

Figure A.1: KT2 TCR sequences
The proportion of sequenced KT2 TCR molecules (α chain in (a), β chain in (b)) categorised by Decombinator as each of the identified TCR clonotypes. The largest identified clonotypes are coloured and described by the five-part Decombinator identifier (V region, J region, V deletions, J deletions, junctional nucleotides).

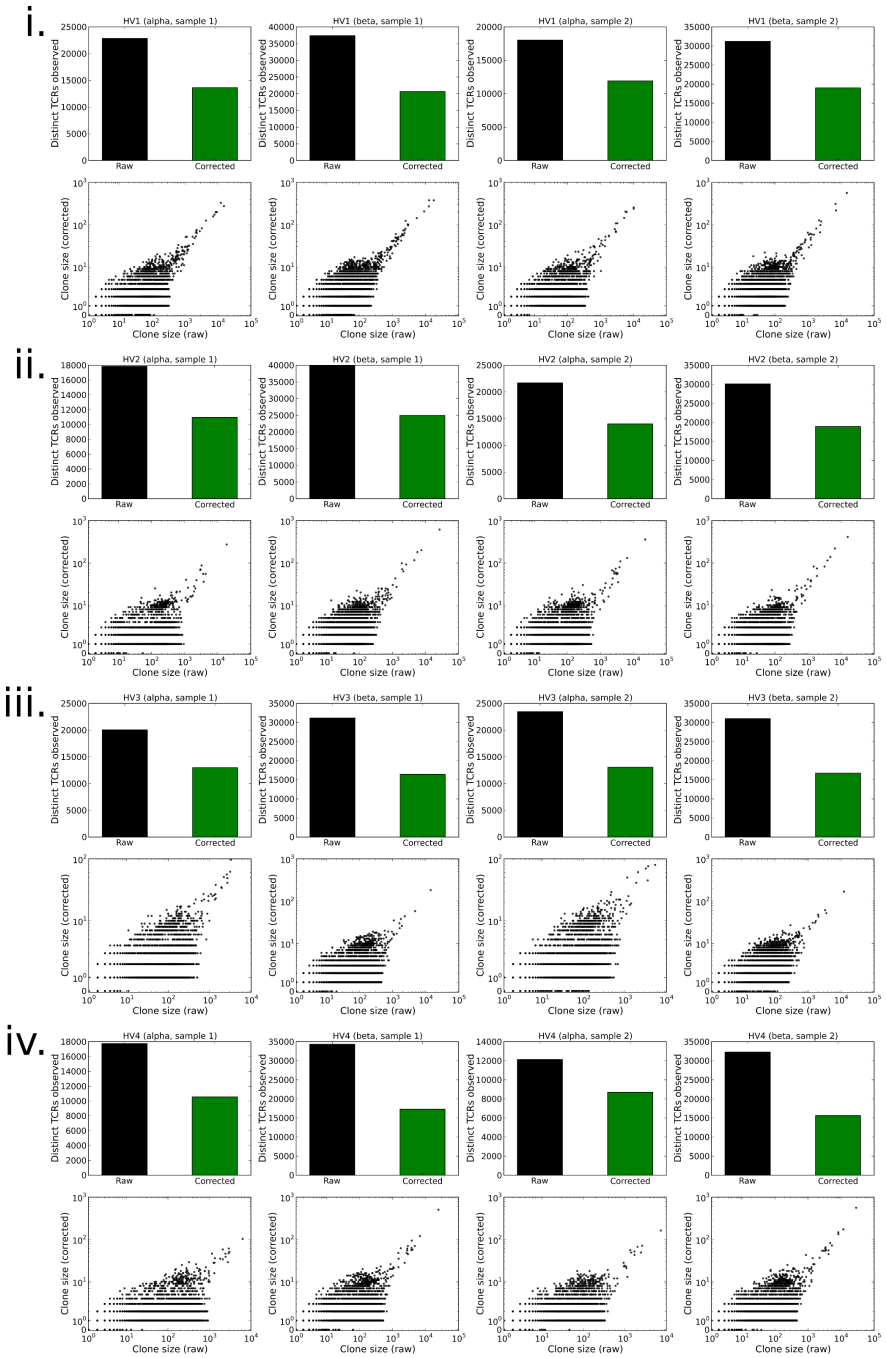

Figure A.2: Effect of error correction on observed TCR repertoire

The effect of the correction of sequencing error and biased PCR amplification on sequencing runs of TCRs from healthy volunteers for alpha and beta samples at two different time points from 4 healthy volunteers (Figures i-iv).

Top panels: The number of distinct TCR clones observed in the indicated sequencing run of healthy volunteer peripheral blood ('PB') when barcodes are not considered ('raw') and when barcodes are used to correct for PCR amplification bias and sequencing error ('corrected').

Bottom panels: The correlation between TCR clone size observed in the raw or corrected data of the indicated sequencing run of healthy volunteer PB.

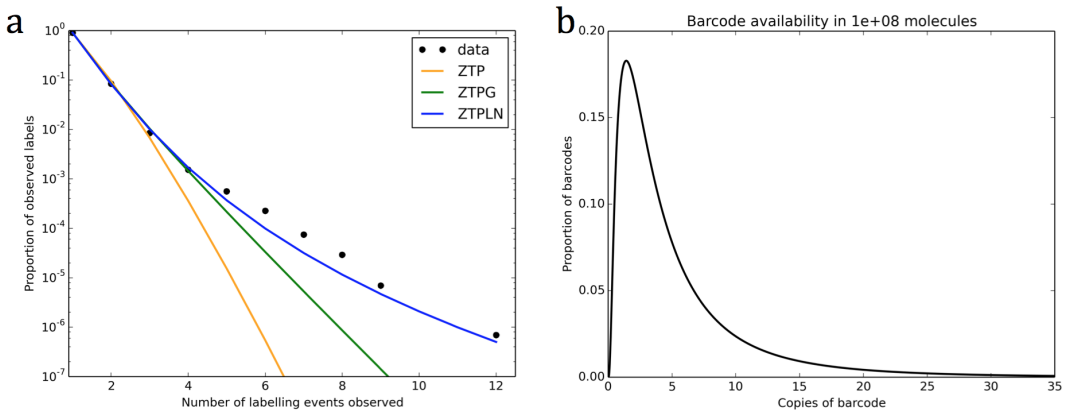

Figure A.3: Distribution of available barcode oligonucleotides

(a) Black circles: The proportion of barcodes that are observed with the given number of labelling events across all healthy volunteer PBMC sequencing runs in this study. Coloured lines: fitted mixed Poisson models, zero-truncated to account for unobservable zeros. The distributions considered are the zero-truncated Poisson (ZTP), the zero-truncated mixed Poisson-Gamma (ZTPG), and the zero-truncated mixed Poisson-Lognormal (ZTPLN). Zero-truncation occurs after mixing, and best fitting parameters are found via maximum likelihood estimation.
(b) From the parameters of the best-fitting ZTPLN in (a), the inferred structure of the available barcode pool, giving the distribution of barcode copy numbers in a pool of 10^8 oligos.

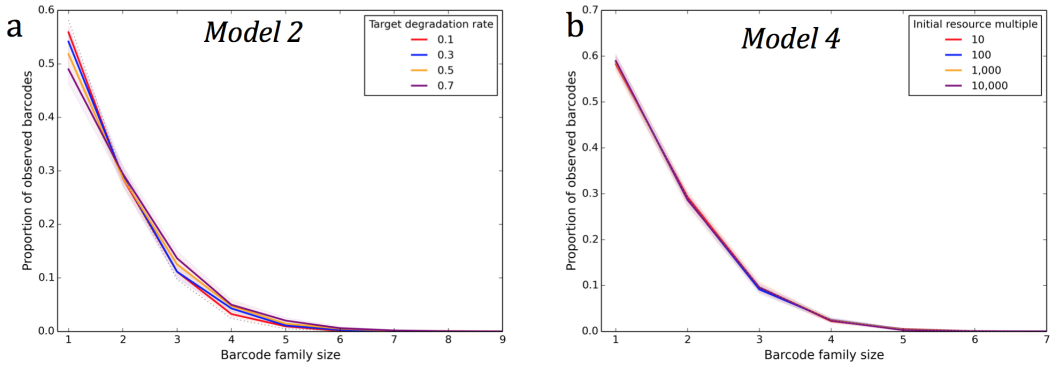

Figure A.4: PCR simulator - models 2 and 4

Observed barcode family size distributions observed under different models of PCR duplication. Simulations performed with 10,000 initial molecules, 25 cycles of PCR (with no error) and sequencing of 10,000 molecules selected from the amplified pool. Simulations were repeated 10 times and the mean and standard deviation are shown. The dotted lines represent the expected distribution if every initial molecule is labelled uniquely and represented equally in the amplified pool.
(a) Model 2: PCR cycles with target degradation. In each cycle, a molecule replicates with probability 0.8. If successful replication does not occur, degradation of that target molecule occurs with the indicated probability.
(b) Model 4: Resource degradation model of PCR. An initial amount of abstract resource is available at the start of the process, given as a multiple (the initial resource multiple) of the number of initial molecules. The efficiency of the reaction in a given cycle depends on the amount of resource available. A successful replication depletes resource at a fixed rate, here 0.5.

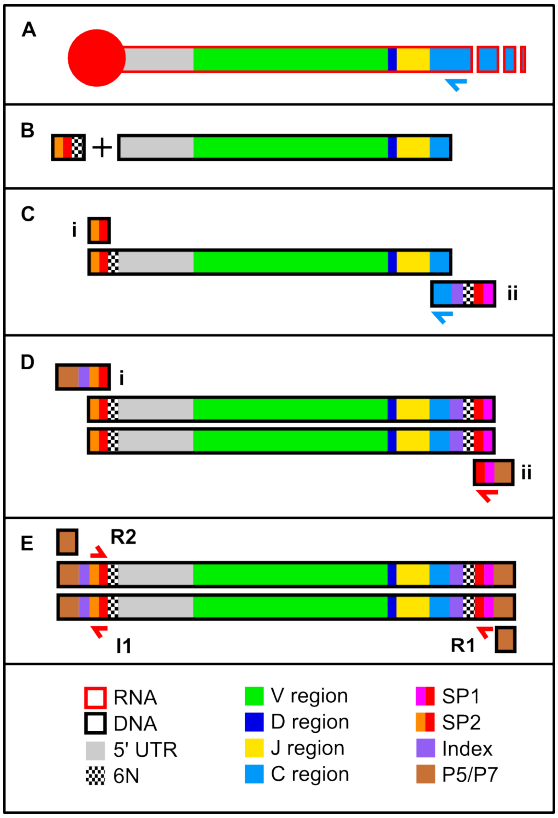

Figure A.5: Schematic of Protocol A (using single strand ligation)

Figure producd by James Heather

A: TCR RNA is reverse transcribed using oligonucleotides directed against the 5' of the alpha and beta constant regions (blue arrow). Red circle represents mRNA 5' cap.

B: RACE is achieved through ligation of DNA adapter to the 3' end of the cDNA; adapter consists of Illumina sequencing primer SP2 and a hexamer of random nucleotides (6N).

C: Separate single round second (i) and third strand (ii) reactions allow incorporation of another random hexamer at the other end of the amplicons (as well as SP1, the other sequencing primer, and an index for demultiplexing), completing the unique 12-mer barcoding of TCR cDNA.

D: A four-cycle PCR is used to add: another index (that which is sequenced in the dedicated Illumina indexing read) and P7 at one end (i) and P5 at the other (ii). P5 and P7 are the elements required for cluster generation, as they bind to the oligonucleotides that coat the flow-cell, permitting bridge amplification.

E: A final PCR directed against the P5 and P7 elements (for 23 cycles) amplifies full-length amplicons to sufficient concentrations for sequencing. Amplicons are finally purified, quantified, sized and normalised before sequencing on the MiSeq.

Appendix B

PCR simulator code

Below is the Python code used for the models of PCR amplification presented in Chapter 2.

```
############################################################################################################
# PCRsim v 1.1
#
# Functions simulating PCR amplification of cDNA with unique molecular labelling and high
# throughput sequencing
#
# Run as a script to see output of example usage in the
# "if __name__ == __main__" section at the bottom
############################################################################################################

from __future__ import division
import collections as coll
import random
import itertools
import numpy as np

############################################################################################################
# 1. Create intial pool of molecules

def create_initial_molecules(
  number_initial_molecules, nt_string=False, length=10, identical=True):
    """
    Creates a dictionary representing the initial sample to be used in further steps of the
    simulator
    Returns a dictionary in the form {template_molecule_identifier: number_of_copies}

     - if nt_string is True, template_molecule_identifier will be a nucleotide string, of specified
       length. This is created randomly.

     - if identical is False, a different template_molecule_identifier string will be created for
       each of the molecules in the initial sample, representing a polyclonal sample.

    Note that if nt_string is False, error cannot be simulated in later steps.
    """

    if nt_string == False:
        return coll.Counter({'target': number_initial_molecules})

    if nt_string == True:
```

```
        if identical == True:
            template = ''.join([random.choice(['A', 'C', 'G', 'T']) for _ in range(length)])
            return coll.Counter({template: number_initial_molecules})
        if identical == False:
            molecule_dictionary = coll.Counter()
            for n in xrange(number_initial_molecules):
                molecule = ''.join([random.choice(['A', 'C', 'G', 'T']) for _ in range(length)])
                molecule_dictionary[molecule] += 1
            return molecule_dictionary

##############################################################################################
# 2. Label initial pool of molecules

def label_initial_molecules(
  initial_pool, unique=True, nt_string=False, label_length = 6,
  number_labels_available=10**6, label_distribution='uniform'):
    """
    Takes a dictionary representing the initial sample
    Returns a dictionary representing the intial sample with ligated identifying barcodes

     - If unique is True, each initial molecule is guaranteed to be uniquely labelled. If unique is
       set to False, the labels are chosen from the available pool randomly and there is a chance
       that the same label is ligated to more than one molecule.

     - The format of the label is determined by the nt_string argument:
        - If nt_string is False, the molecules are labelled by a number, and the label_length
          argument is not used.
        - If nt_string is True, the molecules are labelled by a nucletoide string and the
          number_labels_available argument is not used.

     - If nt_string is False and unique is False, the label number is selected at random from the
       integers between 1 and number_labels_available.

     - If nt_string is True and unique is False a random nucleotide string of length label_length is
       ligated to each initial molecule.

     - The label_distribution argument describes the structure of the pool of labels that is being
       selected from to barcode each molecule. The options are normal or lognormal, and should be
       given in the format "distribution,parameter1,parameter2" (eg "normal,5,2" for a normally
       distributed barcode pool with mean 5 and standard deviation 2)

    Note that if unique=True is used in conjunction with nt_string=True the assigned labels are
    taken sequentially from the possible labels, meaning that the labels are 'closer' to each other
    (in terms of string distance) that if they were chosen at random. This may have an impact on the
    effect that PCR or sequencing error has on the output.
    """

    labelled_pool = coll.Counter()

    # Unique = True overrides barcode_distribution option
    if unique == True and nt_string == False:
        count = 0
        for k in initial_pool.keys():
            for molecule in range(initial_pool[k]):
                labelled_pool[count] += 1
                count += 1
        return labelled_pool

    if unique == True and nt_string == True:
        labels = itertools.imap(''.join, itertools.product('ACGT', repeat=label_length))
        for k in initial_pool.keys():
            for molecule in range(initial_pool[k]):
                mol_str = labels.next()+k
                labelled_pool[mol_str] += 1
        return labelled_pool
```

```
if label_distribution.lower().startswith('u'):

    # Uniformly distributed available barcodes

    if unique == False and nt_string == False and label_distribution == 'uniform':
        labels = xrange(number_labels_available)
        for k in initial_pool.keys():
            for molecule in range(initial_pool[k]):
                mol_str = str(random.choice(labels))+'_'+k
                labelled_pool[mol_str] += 1
        return labelled_pool

    if unique == False and nt_string == True and label_distribution == 'uniform':
        for k in initial_pool.keys():
            for molecule in range(initial_pool[k]):
                label = ''.join([random.choice(['A', 'C', 'G', 'T']) for _ in range(label_length)])
                mol_str = label + k
                labelled_pool[mol_str] += 1
        return labelled_pool

if label_distribution.lower().startswith('n'):

    # Normally distributed available barcodes.

    distribution_mean = float(label_distribution.split(',')[1])
    distribution_sd = float(label_distribution.split(',')[2])

    if not nt_string:
        barcode_weights = {x:max(random.gauss(distribution_mean, distribution_sd),0) for x \
            in xrange(number_labels_available)}
        for k in initial_pool.keys():
            for molecules in range(initial_pool[k]):
                mol_label = weighted_sample_with_replacement_from_counter(barcode_weights, 1).keys()[0]
                mol_str = str(mol_label) + '_' + k
                labelled_pool[mol_str] += 1
        return labelled_pool

    if nt_string:
        barcode_weights = coll.defaultdict(float)
        for bc in itertools.product('ACGT', repeat=label_length):
            barcode_weights[''.join(bc)] = max(random.gauss(distribution_mean, distribution_sd), 0)
        for k in initial_pool.keys():
            for molecules in range(initial_pool[k]):
                mol_label = weighted_sample_with_replacement_from_counter(barcode_weights, 1).keys()[0]
                mol_str = ''.join([mol_label, k])
                labelled_pool[mol_str] += 1
        return labelled_pool

if label_distribution.lower().startswith('l'):

    # Lognormally distributed available barcode

    distribution_lmu = float(label_distribution.split(',')[1])
    distribution_lsd = float(label_distribution.split(',')[2])

    if not nt_string:
        barcode_weights = {x:random.lognormvariate(distribution_lmu, distribution_lsd) for x \
            in xrange(number_labels_available)}
        for k in initial_pool.keys():
            for molecules in range(initial_pool[k]):
                mol_label = weighted_sample_with_replacement_from_counter(barcode_weights, 1).keys()[0]
                mol_str = str(mol_label) + '_' + k
                labelled_pool[mol_str] += 1
        return labelled_pool
```

```
        if nt_string:
            barcode_weights = coll.defaultdict(float)
            for bc in itertools.product('ACGT', repeat=barcode_length):
                barcode_weights[''.join(bc)] = random.lognormvariate(distribution_lmu, distribution_lsd)
            for k in initial_pool.keys():
                for molecules in range(initial_pool[k]):
                    mol_label = weighted_sample_with_replacement_from_counter(barcode_weights, 1).keys()[0]
                    mol_str = ''.join([mol_label, k])
                    labelled_pool[mol_str] += 1
            return labelled_pool


##############################################################################################
# 3. PCR models

def pcr_model1(molecules, error=0, number_cycles=10, efficiency=0.9):
    """
    Takes a dictionary representing the initial sample of molecules.
    Returns a dictionary representing the amplified sample of molecules.

    PCR model 1 is a straightforward branching process model where in every cycle every molecule has
    the same chance of being replicated once, given by the efficiency argument.

    If error > 0, at every cycle the molecule being produced has a chance of having error
    incorporated, given by the error argument (chance of error per base).
    """

    d = molecules.copy()

    if error == 0:
        for c in range(number_cycles):
            for k, v in d.iteritems():
                number_successfully_replicated = np.random.binomial(v, efficiency)
                d[k] += number_successfully_replicated
        return d

    elif error > 0:
        for c in range(number_cycles):
            print 'cycle '+str(c+1)+' ('+"{0:,}".format(sum(d.values()))+' molecules)'
            new_d = d.copy()
            for k, v in d.iteritems():
                for _ in range(v):
                    successful_replication = np.random.binomial(1, efficiency)
                    if successful_replication:
                        new_string = create_error(k, error)
                        new_d[new_string] += 1
            d = new_d.copy()
        return d

def pcr_model2(molecules, error=0, number_cycles=10, efficiency=0.9, degradation=0.1):
    """
    Takes a dictionary representing the initial sample of molecules.
    Returns a dictionary representing the amplified sample of molecules.

    PCR model 2 is a target degradation model. In every cycle every molecule has the same chance of
    being replicated once, given by the efficiency argument. If a molecule does not successfully
    replicate then it might degrade, with probability given by the degradation argument.

    If error > 0, at every cycle the molecule being produced has a chance of having error
    incorporated, given by the error argument (chance of error per base).
    """

    d = molecules.copy()
```

```
    if error == 0:
        for c in range(number_cycles):
            for k, v in d.iteritems():
                if v > 0:
                    number_successfully_replicated = np.random.binomial(v, efficiency)
                    number_not_replicated = v - number_successfully_replicated
                    if number_not_replicated:
                        number_degraded = np.random.binomial(number_not_replicated, degradation)
                    else:
                        number_degraded = 0
                    d[k] += (number_successfully_replicated - number_degraded)
        return d

    elif error > 0:
        for c in range(number_cycles):
            print 'cycle '+str(c+1)+' ('+"{0:,}".format(sum(d.values()))+' molecules)'
            new_d = d.copy()
            for k, v in d.iteritems():
                for _ in range(v):
                    successful_replication = np.random.binomial(1, efficiency)
                    if successful_replication:
                        new_string = create_error(k, error)
                        new_d[new_string] += 1
                    else:
                        molecule_degrades = np.random.binomial(1, degradation)
                        if molecule_degrades:
                            new_d[k] -= 1
            d = new_d.copy()
        return d

def pcr_model3(molecules, error=0, number_cycles=10, available_resource_multiple=50):
    """
    Takes a dictionary representing the initial sample of molecules.
    Returns a dictionary representing the amplified sample of molecules.

    PCR model 3 is a competition model. In every cycle every molecule has the same chance of being
    replicated once. This chance is determined by the amount of abstract "resource" available, which
    is constant throughout the experiment as given as a multiple of the number of initial molecules
    (the available_resource_multiple argument). The chance of successful replication in a given
    cycle is calculated as the available resource divided by  the number of molecules present at the
    start of the cycle (restricted to 0 - 1).

    If error > 0, at every cycle the molecule being produced has a chance of having error
    incorporated, given by the error argument (chance of error per base).
    """

    d = molecules.copy()
    available_resource = available_resource_multiple * sum(d.values())

    if error == 0:
        for c in range(number_cycles):
            cycle_efficiency = min(available_resource/sum(d.values()), 1)
            for k, v in d.iteritems():
                number_successfully_replicated = np.random.binomial(v, cycle_efficiency)
                d[k] += number_successfully_replicated
        return d

    elif error > 0:
        for c in range(number_cycles):
            print 'cycle '+str(c+1)+' ('+"{0:,}".format(sum(d.values()))+' molecules)'
            new_d = d.copy()
            cycle_efficiency = min(available_resource/sum(new_d.values()), 1)
            for k, v in d.iteritems():
                for _ in range(v):
                    successful_replication = np.random.binomial(1, cycle_efficiency)
```

```
                    if successful_replication:
                        new_string = create_error(k, error)
                        new_d[new_string] += 1
            d = new_d.copy()
        return d


def pcr_model4(
  molecules, error=0, number_cycles=10, initial_resource_multiple=50, resource_degradation=0.2):
    """
    Takes a dictionary representing the initial sample of molecules.
    Returns a dictionary representing the amplified sample of molecules.

    PCR model 4 is a competition model, incorporating resource degradation. In every cycle every
    molecule has the same chance of being replicated once. This chance is determined by the amount
    of abstract "resource" available, which is degraded as the reaction progresses. The initial
    amount of available resource is given by the number of initial molecules multipled by the
    initial_resource_multiple argument. Every successful replication reduces the available resource
    by an amount given by the resource_degradation paramter. The chance of successful replication in
    a cycle is given by the amount of resource available at the start of the cycle divided by the
    number of molecules present at the start of the cycle (restricted to 0 - 1).

    If error > 0, at every cycle the molecule being produced has a chance of having error
    incorporated, given by the error argument (chance of error per base).
    """

    d = molecules.copy()
    current_resource = initial_resource_multiple * sum(d.values())

    if error == 0:
        for c in range(number_cycles):
            cycle_efficiency = min(current_resource/sum(d.values()), 1)
            for k, v in d.iteritems():
                number_successfully_replicated = np.random.binomial(v, cycle_efficiency)
                d[k] += number_successfully_replicated
                current_resource -= (number_successfully_replicated * resource_degradation)
                current_resource = max(current_resource, 0)

        return d

    elif error > 0:
        for c in range(number_cycles):
            print 'cycle '+str(c+1)+' ('+"{0:,}".format(sum(d.values()))+' molecules)'
            new_d = d.copy()
            cycle_efficiency = min(current_resource/sum(d.values()), 1)
            for k, v in d.iteritems():
                number_successfully_replicated = 0
                for _ in range(v):
                    successful_replication = np.random.binomial(1, cycle_efficiency)
                    if successful_replication:
                        number_successfully_replicated += 1
                        new_string = create_error(k, error)
                        new_d[new_string] += 1
                current_resource -= (number_successfully_replicated * resource_degradation)
                current_resource = max(current_resource, 0)
            d = new_d.copy()

        return d


def pcr_model5(
  molecules, error=0, number_cycles=10, efficiency_distribution='normal',
  efficiency_parameters=(0.8,0.1)):
    """
    Takes a dictionary representing the initial sample of molecules.
```

```
    Returns a dictionary representing the amplified sample of molecules.

    PCR model 5 is a variable efficiency model. In every cycle, the probability that a molecule will
    successfully replicate is chosen from a distribution of efficiencies. The arguments
    efficiency_distribution and efficiency_parameters determine the distribution that the efficiency
    for one molecule in one cycle is selected from.

    Valid efficiency distribution families are normal and lognormal. If a value < 0 or > 1 is chosen
    from the specified distribution this is replaced with 0 or 1 as appropriate.

    If error > 0, at every cycle the molecule being produced has a chance of having error
    incorporated, given by the error argument (chance of error per base).
    """

    d = molecules.copy()

    avg_effs = list()

    if error == 0:
        for c in range(number_cycles):
            print 'cycle '+str(c+1)+' ('+"{0:,}".format(sum(d.values()))+' molecules)'
            for k, v in d.iteritems():
                efficiencies = pick_efficiencies(efficiency_distribution, efficiency_parameters, v)
                avg_effs.append(np.mean(efficiencies))
                number_successfully_replicated = 0
                while efficiencies:
                    number_successfully_replicated += np.random.binomial(1, efficiencies.pop(0))
                d[k] += number_successfully_replicated

        print 'average efficiency ', np.mean(avg_effs)
        return d

    elif error > 0:
        for c in range(number_cycles):
            print 'cycle '+str(c+1)+' ('+"{0:,}".format(sum(d.values()))+' molecules)'
            new_d = d.copy()
            for k, v in d.iteritems():
                efficiencies = pick_efficiencies(efficiency_distribution, efficiency_parameters, v)
                avg_effs.append(np.mean(efficiencies))
                for i, _ in enumerate(range(v)):
                    successful_replication = np.random.binomial(1, efficiencies[i])
                    if successful_replication:
                        new_string = create_error(k, error)
                        new_d[new_string] += 1
            d = new_d.copy()

        print 'average efficiency ', np.mean(avg_effs)
        return d

def pcr_model6(
  molecules, error=0, number_cycles=10, efficiency_distribution='normal',
  efficiency_parameters=(0.8,0.1), eff_seq_dependent=False):
    """
    Takes a dictionary representing the initial sample of molecules.
    Returns a dictionary representing the amplified sample of molecules.

    PCR model 6 is a variable inherited efficiency model. In every cycle, the probability that a
    molecule will successfully replicate is inherited from its ancestors. Before the PCR reaction
    begins, each initial molecule is assigned an efficiency from the efficiency distribution defined
    by the efficiency_distribution and efficiency_parameters arguments. This is then the probability
    that this molecule and all its descendants successfully replicate in any given cycle.

    Valid efficiency distribution families are normal and lognormal. If a value < 0 or > 1 is chosen
    from the specified distribution this is replaced with 0 or 1 as appropriate.
```

```
    If error > 0, at every cycle the molecule being produced has a chance of having error
    incorporated, given by the error argument (chance of error per base).

    When PCR error occurs, the eff_seq_dependent argument determines whether the new molecule (which
    has a different nucleotide sequence from its ancestors) continues to have the same probability
    of replication. If eff_seq_dependent is True, then a new efficiency is selected whenever a
    different nucleotide sequence is encountered.
    """

    d = molecules.copy()
    efficiencies = coll.defaultdict(float, {k:pick_efficiencies(efficiency_distribution,
                                            efficiency_parameters, 1)[0] for k in d.keys()})

    if error == 0:
        for c in range(number_cycles):
            for k, v in d.iteritems():
                number_successfully_replicated = np.random.binomial(v, efficiencies[k])
                d[k] += number_successfully_replicated
        return d

    elif error > 0:
        for c in range(number_cycles):
            print 'cycle '+str(c+1)+' ('+"{0:,}".format(sum(d.values()))+' molecules)'
            new_d = d.copy()
            for k, v in d.iteritems():
                for _ in range(v):
                    successful_replication = np.random.binomial(1, efficiencies[k])
                    if successful_replication:
                        new_string = create_error(k, error)
                        new_d[new_string] += 1
                        if not eff_seq_dependent:
                            efficiencies[new_string] = efficiencies[k]
                        if eff_seq_dependent:
                            if not efficiencies[new_string]:
                                efficiencies[new_string] = pick_efficiencies(efficiency_distribution,
                                                                             efficiency_parameters, 1)[0]
            d = new_d.copy()

        return d

def create_error(nt_string, error_rate):
    """
    Introduces error at the specified error_rate (rate per base) to the provided nucleotide string.
    Returns the new nucleotide string with error (if any introduced).
    """
    original_string_split = list(nt_string)
    new_string_split = ['']*len(original_string_split)
    for pos, nt in enumerate(original_string_split):
        error_here = np.random.binomial(1, error_rate)
        if error_here:
            nts = ['A', 'C', 'G', 'T']
            nts.remove(nt)
            new_nt = random.choice(nts)
            new_string_split[pos] = new_nt
        else:
            new_string_split[pos] = nt
    return ''.join(new_string_split)

def appropriate_efficiency(a):
    """
    Returns a number between 0 and 1 inclusive.
    """
    a = max(a, 0)
    a = min(a, 1)
    return a
```

```
def pick_efficiencies(efficiency_distribution, efficiency_parameters, n):
    """
    Selects replication efficiencies from the specified distribution
    """
    if str.lower(efficiency_distribution) in ['normal', 'norm', 'n', 'gaussian', 'gauss']:
        return [appropriate_efficiency(random.gauss(efficiency_parameters[0],
                                                    efficiency_parameters[1])) for _ in xrange(n)]

    if str.lower(efficiency_distribution) in ['lognormal', 'lognorm', 'ln']:
        return [appropriate_efficiency(random.lognormvariate(efficiency_parameters[0],
                                                     efficiency_parameters[1])) for _ in xrange(n)]

################################################################################################
# 4. Sampling function

def weighted_sample_no_replacement_from_counter(counter, sample_size):
    """
    Takes a dictionary representing a pool of molecules.
    Returns a dictionary representing a sample selected from the provided pool.
    """

    keys_weights_tuples = [(k, v) for k, v in counter.iteritems()]
    sample = coll.Counter()
    totals = np.cumsum([p[1] for p in keys_weights_tuples])

    for i in xrange(sample_size):
        rnd = random.random() * totals[-1]
        idx = np.searchsorted(totals, rnd, "right")
        sample[keys_weights_tuples[idx][0]] += 1
        totals[idx:] -= 1

    return sample

def weighted_sample_with_replacement_from_counter(counter, sample_size):

    keys_weights_tuples = [(k, v) for k, v in counter.iteritems()]
    sample = coll.Counter()
    totals = np.cumsum([p[1] for p in keys_weights_tuples])

    for i in xrange(sample_size):
        rnd = random.random() * totals[-1]
        idx = np.searchsorted(totals, rnd, "right")
        sample[keys_weights_tuples[idx][0]] += 1

    return sample

################################################################################################
# 5. Sequencing

def sequence_amplified_pool(amplified_pool, sample_size=10**5, error_rate=10**-3):
    """
    Takes a dictionary representing the amplified pool.
    Returns a dictionary representing the sequenced molecules.

    Selects a sample of the given sample_size from the amplified_pool, using the sampling function
    above. For each molecule in the sample, applies the given per base error_rate to simulated
    sequencing error.

    """

    sample = weighted_sample_no_replacement_from_counter(amplified_pool, sample_size)

    if error_rate == 0:
        return sample
```

```
    else:
        sequenced_sample = coll.Counter()
        for k, v in sample.iteritems():
            for _ in range(v):
                new_string = create_error(k, error_rate)
                sequenced_sample[new_string] += 1
        return sequenced_sample


##############################################################################################
# Example usage:

if __name__ == '__main__':

    number_initial_molecules = 1000
    target_length = 25
    barcode_length = 6
    number_pcr_cycles = 12
    pcr_error_rate = 10**-5
    seq_error_rate = 10**-3
    sequence_depth = 500

    # create molecules, identified by nt string, polyclonal
    initial_molecules = create_initial_molecules(number_initial_molecules, nt_string=True,
        length=target_length, identical=False)

    # label molecules, no guarantee of unique labelling, barcode length 6
    labelled_molecules = label_initial_molecules(initial_molecules, unique=False,
        nt_string=True, label_length=barcode_length)

    # alternatively label molecules again from a non-uniformly (normally) distributed pool of available labels
    labelled_molecules_v2 = label_initial_molecules(initial_molecules, unique=False,
        nt_string=False, number_labels_available=1000, label_distribution="n,10,4")

    # use nt labels, normally distributed
    labelled_molecules_v3 = label_initial_molecules(initial_molecules, unique=False,
        nt_string=True, label_length=4, label_distribution="n,10,4")

    # numerical labels, lognormally distributed
    labelled_molecules_v4 = label_initial_molecules(initial_molecules, unique=False,
        nt_string=False, number_labels_available=1000, label_distribution="l,2,1")

    # nt labels, lognormally distributed
    labelled_molecules_v5 = label_initial_molecules(initial_molecules, unique=False,
        nt_string=True, label_length=5, label_distribution="l,2,1")

    # amplify molecules according to pcr model 1 with no error
    print 'amplifying via model 1 with no PCR error...'
    amplified_molecules_model1 = pcr_model1(labelled_molecules, error=0,
        number_cycles=number_pcr_cycles, efficiency=0.8)
    # sequence a sample of molecules
    print 'sampling...'
    sequenced_molecules_model1 = sequence_amplified_pool(amplified_molecules_model1,
        sample_size=sequence_depth, error_rate=seq_error_rate)

    print 'effect of sequencing error:'
    intersect = len(set(sequenced_molecules_model1.keys()).intersection(set(labelled_molecules.keys())))
    total = len(sequenced_molecules_model1.keys())
    print intersect, 'out of', total, 'sequenced clonotypes were in the initial sample'
    print ''

    # amplify molecules according to pcr model 2 (including degradation of target)
    print 'amplifying via model 2...'
    amplified_molecules_model2 = pcr_model2(labelled_molecules, error=pcr_error_rate,
```

```
    number_cycles=number_pcr_cycles, efficiency=0.8, degradation=0.1)
# sequence a sample of molecules
print 'sampling...'
sequenced_molecules_model2 = sequence_amplified_pool(amplified_molecules_model2,
    sample_size=sequence_depth, error_rate=seq_error_rate)

print 'effect of PCR and sequencing error:'
intersect = len(set(sequenced_molecules_model2.keys()).intersection(set(labelled_molecules.keys())))
total = len(sequenced_molecules_model2.keys())
print intersect, 'out of', total, 'sequenced clonotypes were in the initial sample'
print ''

# amplify molecules according to pcr model 3 (fixed available resource)
print 'amplifying via model 3...'
amplified_molecules_model3 = pcr_model3(labelled_molecules, error=pcr_error_rate,
    number_cycles=number_pcr_cycles, available_resource_multiple=1.5**number_pcr_cycles)
# sequence a sample of molecules
print 'sampling...'
sequenced_molecules_model3 = sequence_amplified_pool(amplified_molecules_model3,
    sample_size=sequence_depth, error_rate=seq_error_rate)

print 'effect of PCR and sequencing error:'
intersect = len(set(sequenced_molecules_model3.keys()).intersection(set(labelled_molecules.keys())))
total = len(sequenced_molecules_model3.keys())
print intersect, 'out of', total, 'sequenced clonotypes were in the initial sample'
print ''

# amplify molecules according to pcr model 4 (resource degradation)
print 'amplifying via model 4...'
amplified_molecules_model4 = pcr_model4(labelled_molecules, error=pcr_error_rate,
    number_cycles=number_pcr_cycles, initial_resource_multiple=1000,
    resource_degradation=0.5)
# sequence a sample of molecules
print 'sampling...'
sequenced_molecules_model4 = sequence_amplified_pool(amplified_molecules_model4,
    sample_size=sequence_depth, error_rate=seq_error_rate)

print 'effect of PCR and sequencing error:'
intersect = len(set(sequenced_molecules_model4.keys()).intersection(set(labelled_molecules.keys())))
total = len(sequenced_molecules_model4.keys())
print intersect, 'out of', total, 'sequenced clonotypes were in the initial sample'
print ''

# amplify molecules according to pcr model 5 (variable efficiency)
print 'amplifying via model 5...'
amplified_molecules_model5 = pcr_model5(labelled_molecules, error=pcr_error_rate,
    number_cycles=number_pcr_cycles)
# sequence a sample of molecules
print 'sampling...'
sequenced_molecules_model5 = sequence_amplified_pool(amplified_molecules_model5,
    sample_size=sequence_depth, error_rate=seq_error_rate)

print 'effect of PCR and sequencing error:'
intersect = len(set(sequenced_molecules_model5.keys()).intersection(set(labelled_molecules.keys())))
total = len(sequenced_molecules_model5.keys())
print intersect, 'out of', total, 'sequenced clonotypes were in the initial sample'
print ''

# amplify molecules according to pcr model 6 (variable efficiency)
print 'amplifying via model 6...'
amplified_molecules_model6 = pcr_model6(labelled_molecules, error=pcr_error_rate,
    number_cycles=number_pcr_cycles)
# sequence a sample of molecules
print 'sampling...'
sequenced_molecules_model6 = sequence_amplified_pool(amplified_molecules_model6,
```

```
    sample_size=sequence_depth, error_rate=seq_error_rate)

print 'effect of PCR and sequencing error:'
intersect = len(set(sequenced_molecules_model6.keys()).intersection(set(labelled_molecules.keys())))
total = len(sequenced_molecules_model6.keys())
print intersect, 'out of', total, 'sequenced clonotypes were in the initial sample'
print ''
```

Appendix C

Immune tolerance model: Supplementary Information

The following was primarily written by Chris Watkins in support of the work presented in Chapter 3.

The clonotype update equation (Equation 3.4 in Chapter 3) is an example of a type of 'multiplicative weight update algorithm' which has been extensively studied in machine learning and game theory. These algorithms have excellent convergence properties; a recent expository survey of this work is [10], which develops a unified presentation and analysis of many applications of these algorithms.

Our approach is to write the rate of change of clonotype frequencies as minus the gradient of a convex function F on vectors of N clonotype frequencies. F is defined on the positive quadrant. The updates of clonotype abundances in the model then match exactly a multiplicative update algorithm for finding the minimum of a convex function. F has a unique minimum, and we can derive an estimate of the rate at which this minimum is approached.

A formal development is below: the analysis is adapted from [10].

We follow the approach of [10] closely. Our multiplicative weights algorithm (described below) is similar to theirs, but without the weight normalisation step. Our Theorem 3 is adapted from their Theorem 2.4, with the difference that in ours there is

no weight normalisation, and we therefore use generalised KL-divergence, which is a measure of similarity of two positive vectors $\mathbf{q}$ and $\mathbf{x}$:

$$D(\mathbf{q}\|\mathbf{x}) = \sum_i q_i \ln\left(\frac{q_i}{x_i}\right) + \sum_i x_i - \sum_i q_i \tag{C.1}$$

Our final result, Proposition 5 is essentially their Theorem 3.11, with the difference that we take the limit in continuous time.

C.1 Model and update algorithm

Let there be N T-cell clones, and let the number of cells in clone i at time t be denoted x_i^t, and the vector of all N clone abundances at time t is $\mathbf{x}^t = (x_1^t, \ldots, x_N^t)$; sometimes we will speak of clone abundances without mentioning a specific time, and denote the counts by $\mathbf{x} = (x_1, \ldots, x_N)$. Although in reality the clone sizes would be positive integers, we do not consider small population size effects here, and we model the clone abundances as positive real numbers.

The clone count update algorithm is as follows. Let $\mathbb{R}_+ = \{x \in \mathbb{R} \mid x > 0\}$.

Multiplicative Weights Update Algorithm

The update algorithm for clonotypes is presented below as a sequence of deterministic update steps at discrete times. At each time, the 'signal' for each clonotype is the sum of divide (negative) and die (positive) signals received by cells of that type: for the proof below, these signals are assumed in the range $[-1,1]$, but this normalisation becomes automatic when we take the limit in continuous time.

Initialisation: Fix $\eta \leq \frac{1}{2}$. Let $\mathbf{x}^1 = (x_1^1, \ldots, x_N^1) \in \mathbb{R}_+^N$.
for $t = 1, 2, \ldots, T$:

1. Let $\mathbf{m}^t = (m_1^t, \ldots, m_N^t)$ be the signals received by each of the N clonotypes; these signals are always in the range $[-1,1]$.
2. For all i, $x_i^{t+1} = x_i^t(1 - \eta m_i^t)$

end

Lemma 1 *Let* $0 < \eta \leq \frac{1}{2}$.

For $-1 \leq m < 0$,

$$\ln \frac{1}{1-\eta m} \leq m \ln(1+\eta) \tag{C.2}$$

For $0 \leq m \leq 1$,

$$\ln \frac{1}{1-\eta m} \leq m \ln \frac{1}{1-\eta} \tag{C.3}$$

Proof Case $-1 \leq m < 0$: By concavity of ln, $|m| \ln(1+\eta) \leq \ln(1+|m|\eta)$. Remembering m is negative, it follows that:

$$\begin{aligned} m \ln(1+\eta) &\geq -\ln(1+|m|\eta) \\ &= \ln \frac{1}{1-m\eta} \end{aligned}$$

Case $0 \leq m \leq 1$: Observe that $\ln \frac{1}{1-\eta}$ is convex in η, since:

$$\frac{d^2}{d\eta^2} \ln \frac{1}{1-\eta} = \frac{1}{(1-\eta)^2} > 0$$

Since $\ln \frac{1}{1-\eta} = 0$ when $\eta = 0$, it follows from convexity of $\ln \frac{1}{1-\eta}$ that $\frac{1}{1-m\eta} \leq m \ln \frac{1}{1-\eta}$. ■

Lemma 2 *For* $0 < \eta \leq \frac{1}{2}$,

$$\ln \frac{1}{1-\eta} \leq \eta + \eta^2 \tag{C.4}$$

and

$$\ln(1+\eta) \geq \eta - \eta^2 \tag{C.5}$$

Proof To prove (C.4), consider the derivatives of $u(\eta) = \ln \frac{1}{1-\eta}$ and $v(\eta) = \eta + \eta^2$:

$$u(\eta) = \ln \frac{1}{1-\eta} \qquad v(\eta) = \eta + \eta^2$$
$$u'(\eta) = \frac{1}{1-\eta} \qquad v'(\eta) = 1 + 2\eta$$
$$u''(\eta) = \frac{1}{(1-\eta)^2} \qquad v''(\eta) = 2$$

Observe that $u(0) = v(0)$ and $u'(0) = v'(0)$, but for small η, $u'(\eta) < v'(\eta)$, hence $u(\eta) \leq v(\eta)$ over some interval $[0, \eta^*]$, where $u(\eta^*) = v(\eta^*)$; we need to show that $\frac{1}{2} \leq \eta^*$. At $\eta = \eta^*$, the graph of $u(\eta)$ crosses that of $v(\eta)$ from below, hence $u'(\eta^*) \geq v'(\eta^*)$. Both u' and v' are monotonically increasing; observe that $u'(\eta) = v'(\eta)$ only when $\eta = 0$ or $\eta = \frac{1}{2}$, where $u'(\frac{1}{2}) = v'(\frac{1}{2}) = 2$. It follows that $\eta^* \geq \frac{1}{2}$, and $u(\eta) \leq v(\eta)$ for $\eta \in [0, \frac{1}{2}]$ as required.

For (C.5), consider the derivatives of $w(\eta) = \ln(1+\eta)$ and $y(\eta) = \eta - \eta^2$.

$$w(\eta) = \ln(1+\eta) \qquad y(\eta) = \eta - \eta^2$$
$$w'(\eta) = \frac{1}{1+\eta} \qquad y'(\eta) = 1 - 2\eta$$
$$w''(\eta) = \frac{-1}{(1+\eta)^2} \qquad y''(\eta) = -2$$

Observe that $w(0) = y(0)$, $w'(0) = y'(0)$, and $w''(\eta) > y''(\eta)$ for all $\eta > 0$. Hence $w(\eta) > y(\eta)$ for all $\eta > 0$, which includes what was to be proved. ■

Theorem 3 *Assume that all costs $m_i^t \in [-1, 1]$ and $0 < \eta \leq \frac{1}{2}$. Then the multiplicative weights algorithm described above guarantees that after T rounds of learning, producing the sequence of weight vectors $\mathbf{x}^1, \ldots, \mathbf{x}^T$, and for* any *positive vector* $\mathbf{q}$,

$$\frac{1}{T}\sum_{t=1}^{T} \mathbf{m}^t \cdot \mathbf{x}^t \leq \frac{1}{T}\sum_{t=1}^{T} (\mathbf{m}^t + \eta|\mathbf{m}^t|) \cdot \mathbf{q} + \frac{D(\mathbf{q}\|\mathbf{x}^1)}{\eta T} \tag{C.6}$$

Proof

$$
\begin{aligned}
D(\mathbf{q}\|\mathbf{x}^{t+1}) - D(\mathbf{q}\|\mathbf{x}^{t}) &= \sum_i \left(\left(q_i \ln \frac{q_i}{x_i^{t+1}} + x_i^{t+1} - q_i \right) - \left(q_i \ln \frac{q_i}{x_i^{t}} + x_i^{t} - q_i \right) \right) \\
&= \sum_i q_i (\ln x_i^t - \ln x_i^{t+1}) + \sum_i x_i^{t+1} - x_i^t \\
&= \sum_i q_i (\ln x_i^t - \ln x_i^t (1 - \eta m_i^t)) + \sum (1 - \eta m_i^t) x_i^t - x_i^t \\
&= \sum_i q_i \ln \frac{1}{1 - \eta m_i^t} - \eta \sum_i m_i^t x_i^t \\
&= \sum_{i:m_i^t \geq 0} q_i \ln \frac{1}{1 - \eta m_i^t} + \sum_{i:m_i^t < 0} q_i \ln \frac{1}{1 - \eta m_i^t} - \eta \sum_i m_i^t x_i^t
\end{aligned}
$$

using lemma 1, we obtain

$$
\leq \ln \frac{1}{1-\eta} \sum_{i:m_i^t \geq 0} q_i m_i^t + \ln(1+\eta) \sum_{i:m_i^t < 0} q_i m_i^t - \eta \sum_i m_i^t x_i^t
$$

and using lemma 2, we obtain

$$
\begin{aligned}
&\leq (\eta + \eta^2) \sum_{i:m_i^t \geq 0} q_i m_i^t + (\eta - \eta^2) \sum_{i:m_i^t < 0} q_i m_i^t - \eta \sum_i m_i^t x_i^t \\
&= \eta \left(\sum_i (m_i^t + \eta |m_i^t|) q_i - \sum_i m_i^t x_i^t \right) \\
&= \eta \left((\mathbf{m}^t + \eta |\mathbf{m}^t|) \cdot \mathbf{q} - \mathbf{m}^t \cdot \mathbf{x}^t \right)
\end{aligned}
$$

Summing from $t = 1$ to T, we obtain:

$$
D(\mathbf{q}\|\mathbf{x}^{T+1}) - D(\mathbf{q}\|\mathbf{x}^{1}) = \eta \left(\sum_{t=1}^{T} (\mathbf{m}^t + \eta |\mathbf{m}^t|) \cdot \mathbf{q} - \mathbf{m}^t \cdot \mathbf{x}^t \right)
$$

Rearranging, we obtain:

$$
\frac{1}{T} \sum_{t=1}^{T} \mathbf{m}^t \cdot \mathbf{x}^t \leq \frac{1}{T} \sum_{t=1}^{T} (\mathbf{m}^t + \eta |\mathbf{m}^t|) \cdot \mathbf{q} + \frac{D(\mathbf{q}\|\mathbf{x}^1) - D(\mathbf{q}\|\mathbf{x}^{T+1})}{\eta T}
$$

KL divergence is non-negative, so that $D(\mathbf{q}\|\mathbf{x}^{T+1}) \geq 0$; we can therefore add this term to the RHS, obtaining

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{m}^t\cdot\mathbf{x}^t \leq \frac{1}{T}\sum_{t=1}^{T}(\mathbf{m}^t+\eta|\mathbf{m}^t|)\cdot\mathbf{q}+\frac{D(\mathbf{q}\|\mathbf{x}^1)}{\eta T}$$

which is what was to be proved. ■

We envisage the T-cell birth-death process operating in continuous time. To approach continuous time in the limit of a sequence of small time-steps, we re-write this inequality in terms of time-steps of length δ, so that the total number of time-steps becomes $\frac{T}{\delta}$, and within each time-step there is a multiplicative update with factor $\eta\delta$:

$$\frac{1}{T/\delta}\sum_{t=1}^{T/\delta}\mathbf{m}^t\cdot\mathbf{x}^t \leq \frac{1}{T/\delta}\sum_{t=1}^{T/\delta}(\mathbf{m}^t+\delta\eta|\mathbf{m}^t|)\cdot\mathbf{q}+\frac{D(\mathbf{q}\|\mathbf{x}^1)}{\eta T}$$

Using a bar to denote the time-average during the period 1 to T, and letting δ tend to zero, it follows immediately that:

Corollary 4

$$\overline{\mathbf{m}\cdot\mathbf{x}} \leq \overline{\mathbf{m}\cdot\mathbf{q}}+\frac{D(\mathbf{q}\|\mathbf{x}^1)}{\eta T} \tag{C.7}$$

assuming that the averages exist in the limit as δ tends to zero; this will be the case for the benign choices of $\mathbf{m}$ that we make below.

C.2 On-line minimisation of a convex function

As [10] describe, the multiplicative weights algorithm can be applied to on-line minimisation of a convex function, and theorem 3 can be applied to obtain explicit bounds on the average regret. Let F be a differentiable convex function on $\mathbb{R}^N_+$, and let

$$\rho = \max_{\mathbf{x},i}\left|\frac{\partial F(\mathbf{x})}{\partial x_i}\right| \tag{C.8}$$

That is, ρ is the maximum absolute partial gradient of F anywhere in any of the N coordinate directions. In fact it is only necessary to take ρ to be the maximum absolute such gradient that is actually encountered during the optimisation: in our optimisations, and for the F we use, ρ is unproblematically finite. Now define

$$\mathbf{m}^t = \frac{1}{\rho}\nabla F(\mathbf{x}^t) \tag{C.9}$$

so that $m_i^t \in [-1,1]$ for all i and t; since F is assumed convex and differentiable then for all $\mathbf{q}$ and $\mathbf{x}$:

$$F(\mathbf{x}) - F(\mathbf{q}) \leq \nabla F(\mathbf{x}) \cdot (\mathbf{x} - \mathbf{q}) \tag{C.10}$$

Combining this with equation (C.7) we obtain

$$\begin{aligned}
\overline{F(\mathbf{x}) - F(\mathbf{q})} &\leq \overline{\nabla F(\mathbf{x}) \cdot (\mathbf{x} - \mathbf{q})} && \text{(C.11)}\\
&= \overline{\nabla F(\mathbf{x}) \cdot \mathbf{x}} - \overline{\nabla F(\mathbf{x}) \cdot \mathbf{q}} && \text{(C.12)}\\
&= \rho(\overline{\mathbf{m} \cdot \mathbf{x}} - \overline{\mathbf{m} \cdot \mathbf{q}}) && \text{(C.13)}\\
&\leq \frac{D(\mathbf{q}\|\mathbf{x}^1)}{(\eta/\rho)T} && \text{(C.14)}
\end{aligned}$$

For our model, $(\eta/\rho)\frac{\partial F(\mathbf{x})}{\partial x_i}$ is the rate of growth of the count of cells in clonotype i, and we assume that is always negative for sufficiently large $\mathbf{x}$; F must therefore have an infimum in the positive quadrant (including the axis planes, to allow for zeros): let this infimum of F be achieved at $\mathbf{q}$. Then, taking $0 \ln 0 = 0$ so that $D(\mathbf{q}||\mathbf{x}^1)$ exists even if some elements of $\mathbf{q}$ are zero, we have:

Proposition 5

$$\overline{F(\mathbf{x}) - \inf_{\mathbf{q} \in \mathbb{R}_+^N} F(\mathbf{q})} \leq \frac{D(\mathbf{q}\|\mathbf{x}^1)}{(\eta/\rho)T} \tag{C.15}$$

The RHS of the equation above is of the form C/T, where C is a constant, because $D(\mathbf{q}\|\mathbf{x}^1)$, η, and ρ do not vary throughout learning. From the gradient descent argument of equation 10 in the main manuscript, $F(\mathbf{x})$ decreases monotonically during learning; it follows that $F(\mathbf{x}^t)$ declines at least as fast as $\frac{1}{t^2}$. It follows that the constraints are rapidly satisfied since F rapidly approaches its minimum. However, the

number N of clonotypes appears greater than the number of constraints, so that we cannot say how rapidly the clonotype concentrations $\mathbf{x}$ approach $\mathbf{q}$; this may allow clonotype diversity to persist for considerable time, even though the constraints are approximately satisfied.

Appendix D

Short Read Decombinator

D.1 Short Read Decombinator ('SRD') Algorithm

Decombinator [145] is modified for use with short sequence read data as follows.

First, a keyword trie is built for each possible V and J region to be assigned. These tries consist of every contiguous subsequence of at least four nucleotides taken from the 3' portion of a V gene and the 5' portion of a J gene. A Python implementation of the Aho-Corasick algorithm [3] is then used to search the sequence read for all instances of matches from any trie. These matches are used to first determine which V gene should be assigned to the sequence read as follows, and then the same process is followed for J genes.

1. If the longest of all the matches between any gene subsequence and the sequence read is longer than a specified parameter (the 'match threshold') and the longest match to a substring of a different gene is shorter than this longest match by another specified parameter (the 'match differential') then the gene with the longest match is assigned.

2. If the sequence read is unable to be assigned a gene based on the longest match to a gene substring, then the second part of the SRD method is called. Here each gene is given a score of $\sum e^{\text{length}(x)}$ for all matches x between the sequence read and a substring of the gene. An exponentially increasing score function is used

to ensure that a long match carries more weight than two half-length matches.

3. If the score for all gene regions is below a set parameter (the 'score threshold') then the sequence read is discarded as not being similar enough to any gene to justify an assignment. Otherwise, the sequence read is assigned to the highest scoring gene, provided this highest gene score exceeds the next highest by at least an amount defined by a parameter (the 'score differential').

4. If the two highest gene scores are too similar to allow assignment of the gene region that the TCR sequence read contains, the SRD algorithm moves to the final stage where a pairwise alignment is considered between the sequence read and the two highest scoring regions. The sequence read is assigned the gene with the best pairwise alignment to the read.

Once these steps have been implemented for both V and J regions, and a sequence read has been assigned both genes, the number of deletions from each gene is calculated and the nucleotides between the genes are recorded as for standard Decombinator and a 5-part classifier (a 'DCR') fully defining the sequence read is obtained.

The parameters for SRD used in to process the data presented in Chapter 4 are: match threshold = 10, match differential = 2, score threshold = 1400 and score differential = 1.05. These were selected by running a set of artificial sequences (with varying levels of sequencing error) through the method with different sets of parameters. A set was chosen based on a compromise between the accuracy of sequence assignment and the percentage of reads being classified.

D.2 V region pairing

The mouse data analysed in Chapter 4 is obtained through a protocol where V gene specific primers are used and sequence data is obtained from 3' of the primer position. Keyword tries for each V gene for use in the Aho-Corasick search in SRD can then be built based on the sequence of each gene from 3' of the primer site. However, due to the short length of the sequence data and the similarity of some of the V genes at the 3' end, this presents difficulties. The following pairs of genes are unable to be distinguished

by SRD because of their 3' similarity:

Gene name	3' sequence
TRBV24*01	GACTCAGCACTGTACCTCTGTGCCAGCAGTCTGTA
TRBV26*01	GACTCAGCACTGTACCTCTGTGCCAGCAGTCTGTC
TRBV12-1*01	AACTGGAGGACTCTGCTATGTACTTCTGTGCCAGCTCTCTC
TRBV12-2*01	TAGAGGACTCTGCCGTGTACTTCTGTGCCAGCTCTCTC
TRBV13-1*01	TCAGACATCTTTGTACTTCTGTGCCAGCAGTGATG
TRBV13-3*01	CAGACAGCTGTATATTTCTGTGCCAGCAGTGATG

Instead of attempting to distinguish these genes in this work, we instead ignore the second of each pair and assign all matching sequence reads to the first gene.